User Guide

AWS Glue

Copyright © 2025 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

AWS Glue: User Guide

Copyright © 2025 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon's trademarks and trade dress may not be used in connection with any product or service that is not Amazon's, in any manner that is likely to cause confusion among customers, or in any manner that disparages or discredits Amazon. All other trademarks not owned by Amazon are the property of their respective owners, who may or may not be affiliated with, connected to, or sponsored by Amazon.

Table of Contents

What is AWS Glue? ... 1
- AWS Glue features ... 2
- Learning about innovations in AWS Glue ... 3
- Getting started with AWS Glue ... 3
- Accessing AWS Glue ... 4
- Related services ... 4

How it works ... 5
- Serverless ETL jobs run in isolation ... 6
- Concepts ... 7
  - AWS Glue terminology ... 9
- Components ... 12
  - AWS Glue console ... 12
  - AWS Glue Data Catalog ... 12
  - AWS Glue crawlers and classifiers ... 14
  - AWS Glue ETL operations ... 14
  - Streaming ETL in AWS Glue ... 14
  - The AWS Glue jobs system ... 15
  - Visual ETL components ... 15
- AWS Glue for Spark and AWS Glue for Ray ... 21
  - What is AWS Glue for Ray? ... 22
- Converting semi-structured schemas to relational schemas ... 23
- AWS Glue types ... 25
  - AWS Glue Data Catalog Types ... 25
  - Types in AWS Glue with Spark scripts ... 25
  - AWS Glue Crawler Types ... 26

Getting started ... 27
- Overview of using AWS Glue ... 27
- Setting up IAM permissions ... 29
  - Next steps ... 34
  - Setting up for AWS Glue Studio ... 34
  - Getting started with notebooks in AWS Glue Studio ... 45
- Setting up usage profiles ... 48
  - Managing usage profiles ... 49
  - Usage profiles and jobs ... 62

- Getting started with the AWS Glue Data Catalog ... 62
  - Overview ... 63
  - Step 1: Create a database ... 63
  - Step 2. Create a table ... 65
  - Next steps ... 66
- Setting up network access to data stores ... 69
  - Setting up a VPC to connect to PyPI for AWS Glue ... 70
  - Setting up DNS in your VPC ... 72
- Setting up encryption ... 73
- Setting up networking for development ... 76
  - Setting up your network for a development endpoint ... 77
  - Setting up Amazon EC2 for a notebook server ... 79

Data discovery and cataloging ... 80

- Populating the Data Catalog ... 82
  - Using an AWS Glue crawler ... 83
  - Defining metadata manually ... 168
  - Integrating with other AWS services ... 186
  - Data Catalog settings ... 188
- Populating and managing transactional tables ... 191
  - Creating Iceberg tables ... 191
  - Optimizing Iceberg tables ... 195
  - Optimizing query performance for Iceberg tables ... 234
- Managing the Data Catalog ... 236
  - Updating the schema and adding new partitions ... 237
  - Optimizing query performance using column statistics ... 245
  - Encrypting your Data Catalog ... 269
  - Securing your Data Catalog using Lake Formation ... 269
  - Working with AWS Glue Data Catalog views ... 269
- Accessing the Data Catalog ... 274
  - Connecting to the Data Catalog using AWS Glue Iceberg REST endpoint ... 275
  - Connecting to the Data Catalog using AWS Glue Iceberg REST extension endpoint ... 278
  - AWS Glue REST APIs for Apache Iceberg ... 278
  - Connecting to Data Catalog from a standalone Spark application ... 297
  - Data mapping between Amazon Redshift and Apache Iceberg ... 298
  - Considerations and limitations when using AWS Glue Iceberg REST Catalog APIs ... 300
- Data Catalog best practices ... 301

- Monitoring Data Catalog usage metrics in Amazon CloudWatch 301
  - Overview of Data Catalog metrics 302
  - Adding metrics to your CloudWatch dashboard 302
  - 303
- AWS Glue Schema registry 303
  - Schemas 305
  - Registries 307
  - Schema versioning and compatibility 308
  - Open source Serde libraries 313
  - Quotas of the Schema Registry 313
  - How it works 314
  - Getting started 316
  - Integrating with AWS Glue Schema Registry 338
  - Migrating to AWS Glue Schema Registry 364

Connecting to data 366

- Overview of using connectors and connections 366
- Unified connections 367
  - Supported authentication types 371
  - Considerations 371
- Available connections 372
  - Limitations 375
- AWS Glue connection properties 375
  - Required connection properties 376
  - JDBC connection properties 377
  - MongoDB and MongoDB Atlas connection properties 382
  - Salesforce connection properties 383
  - Snowflake connection 383
  - Vertica connection 384
  - SAP HANA connection 385
  - Azure SQL connection 386
  - Teradata Vantage connection 387
  - OpenSearch Service connection 388
  - Azure Cosmos connection 389
  - SSL connection properties 389
  - Kafka connection properties for authentication 392
  - Google BigQuery connection 393

- Vertica connection 384
- Storing connection credentials in AWS Secrets Manager 393
- Adding an AWS Glue connection 394
  - Connecting to Adobe Analytics 396
  - Connecting to Adobe Marketo Engage 404
  - Connecting to Amazon Redshift 414
  - Connecting to Asana 418
  - Connecting to Azure Cosmos DB 428
  - Connecting to Azure SQL 431
  - Connecting to Blackbaud Raiser's Edge NXT 434
  - Connecting to CircleCI 443
  - Connecting to Datadog 455
  - Connecting to Docusign Monitor 462
  - Connecting to Domo 469
  - Connecting to Dynatrace 478
  - Connecting to Facebook Ads 484
  - Connecting to Facebook Page Insights 491
  - Connecting to Freshdesk 513
  - Connecting to Freshsales 522
  - Connecting to Google Ads 529
  - Connecting to Google Analytics 4 556
  - Connecting to Google BigQuery 568
  - Connecting to Google Search Console 572
  - Connecting to Google Sheets 581
  - Connecting to HubSpot 588
  - Connecting to Instagram Ads 614
  - Connecting to Intercom 621
  - Connecting to Jira Cloud 643
  - Connecting to Kustomer 668
  - Connecting to LinkedIn 695
  - Connecting to Mailchimp 703
  - Connecting to Microsoft Dynamics 365 CRM 713
  - Connecting to Microsoft Teams 721
  - Connecting to Mixpanel 731
  - Connecting to Monday 746
  - Connecting to MongoDB 755

- Connecting to Oracle NetSuite ... 758
- Connecting to OpenSearch Service ... 770
- Connecting to Okta ... 773
- Connecting to PayPal ... 781
- Connecting to Pendo ... 788
- Connecting to Pipedrive ... 796
- Connecting to Productboard ... 805
- Connecting to QuickBooks ... 813
- Connecting to Salesforce ... 821
- Connecting to Salesforce Marketing Cloud ... 833
- Connecting to Salesforce Commerce Cloud ... 855
- Connecting to Salesforce Marketing Cloud Account Engagement ... 865
- Connecting to SAP HANA ... 877
- Connecting to SAP OData ... 881
- Connecting to SendGrid ... 902
- Connecting to ServiceNow ... 912
- Connecting to Slack ... 922
- Connecting to Smartsheet ... 933
- Connecting to Snapchat Ads ... 945
- Connecting to Snowflake ... 954
- Connecting to Stripe ... 958
- Connecting to Teradata ... 1006
- Connecting to Twilio ... 1009
- Connecting to Vertica ... 1028
- Connecting to WooCommerce ... 1032
- Connecting to Zendesk ... 1059
- Connecting to Zoho CRM ... 1088
- Connecting to Zoom Meetings ... 1098
- Adding a JDBC connection using your own JDBC drivers ... 1108

Using custom connectors and connections ... 1113
- Creating custom connectors ... 1115
- Creating connections for connectors ... 1119
- Authoring jobs with custom connectors ... 1127
- Managing connectors and connections ... 1134
- Developing custom connectors ... 1137
- Restrictions for using connectors and connections in AWS Glue Studio ... 1139

- Testing an AWS Glue connection 1140
- Configuring AWS calls to go through your VPC 1140
- Connecting to a JDBC data store in a VPC 1141
  - Accessing VPC Data Using elastic network interfaces 1142
  - Elastic network interface properties 1143
- Using a MongoDB or MongoDB Atlas connection 1143
- Crawling an Amazon S3 data store using a VPC endpoint 1144
  - Prerequisites 1144
  - Creating the connection to Amazon S3 1145
  - Testing the connection to Amazon S3 1148
  - Creating a crawler for an Amazon S3 data store 1149
  - Creating a crawler for Amazon S3 backed Data Catalog tables 1152
  - Running a crawler 1153
  - Troubleshooting 1153
- Working with AWS Lake Formation-protected data 1153
  - Full Table Access 1153
- Troubleshooting connection issues 1155
- Tutorial: Using the AWS Glue Connector for Elasticsearch 1156
  - Prerequisites 1156
  - Step 1: (Optional) Create an AWS secret for your OpenSearch cluster information 1157
  - Step 2: Subscribe to the connector 1157
  - Step 3: Activate the connector in AWS Glue Studio and create a connection 1159
  - Step 4: Configure an IAM role for your ETL job 1159
  - Step 5: Create a job that uses the OpenSearch connection 1160
  - Step 6: Run the job 1161

Building AWS Glue jobs with interactive sessions 1162

- Overview of AWS Glue interactive sessions 1162
  - Limitations 1163
- Getting started with AWS Glue interactive sessions 1163
  - Prerequisites for setting up interactive sessions locally 1163
  - Installing Jupyter and AWS Glue interactive sessions Jupyter kernels 1163
  - Running Jupyter 1163
  - Configuring session credentials and region 1164
  - Upgrading from the interactive sessions preview 1165
  - Using interactive sessions with SageMaker AI Studio 1166
  - Using interactive sessions with Microsoft Visual Studio Code 1166

    - Interactive sessions with IAM 1169
  - Configuring AWS Glue interactive sessions for Jupyter and AWS Glue Studio notebooks 1178
    - Introduction to Jupyter Magics 1178
    - Magics supported by AWS Glue interactive sessions for Jupyter 1178
    - Naming sessions 1195
    - Specifying an IAM role for interactive sessions 1196
    - Configuring sessions with named profiles 1196
  - Converting a script or notebook into an AWS Glue job 1197
  - Working with streaming operations in AWS Glue interactive sessions 1198
    - Switching streaming session type 1198
    - Sampling input stream for interactive development 1198
    - Running streaming applications in interactive sessions 1200
  - Developing and testing locally 1201
    - Developing using AWS Glue Studio 1202
    - Developing using interactive sessions 1202
    - Developing AWS Glue jobs locally with Docker 1202
  - Dev endpoints 1212
    - Migrating from dev endpoints to interactive sessions 1213
    - Developing scripts using development endpoints 1215
    - Managing notebooks 1242

Building visual ETL jobs 1244

  - Build visual ETL jobs with AWS Glue Studio 1244
    - Signing in to the console 1244
    - Next steps for creating a job in AWS Glue Studio 1245
  - Build visual ETL flows with Amazon SageMaker 1245
  - Starting visual ETL jobs in AWS Glue Studio 1245
    - Starting jobs in AWS Glue Studio 1246
    - Job editor features 1247
    - Transform data with AWS Glue managed transforms 1255
    - Transform data with custom visual transforms 1328
    - Using Data Lake frameworks with AWS Glue Studio 1345
    - Connecting to data sources 1356
    - Configuring data target nodes 1365
    - Editing or uploading a job script 1370
    - Changing the parent nodes for a node in the job diagram 1374
    - Deleting nodes from the job diagram 1375

  - Adding source and target parameters to the AWS Glue Data Catalog node 1379
  - Using Git version control systems in AWS Glue 1381
- Authoring code with AWS Glue Studio notebooks 1389
  - Limitations 1389
  - Overview of using notebooks 1390
  - Creating an ETL job using notebooks in AWS Glue Studio 1391
  - Notebook editor components 1392
  - Saving your notebook and job script 1393
  - Managing notebook sessions 1394
  - Using Amazon Q Developer with AWS Glue Studio notebooks 1395
- Monitoring job runs 1396
  - Accessing the job monitoring dashboard 1396
  - Overview of the job monitoring dashboard 1396
  - Job runs view 1397
  - Viewing the job run logs 1400
  - Viewing the details of a job run 1401
  - Viewing Amazon CloudWatch metrics for a Spark job run 1404
  - Viewing Amazon CloudWatch metrics for a Ray job run 1404
- Detect and process sensitive data 1406
  - Choosing how you want the data to be scanned 1407
  - Choosing the PII entities to detect 1408
  - Specifying the level of detection sensitivity 1412
  - Choosing what to do with identified PII data 1412
  - Adding fine-grained action overrides 1413
- Managing jobs 1414
  - Start a job run 1415
  - Schedule job runs 1415
  - Manage job schedules 1417
  - Stop job runs 1417
  - View your jobs 1418
  - View information for recent job runs 1418
  - View the job script 1419
  - Modify the job properties 1420
  - Save the job 1422
  - Clone a job 1425
  - Delete jobs 1425

Working with jobs 1427
AWS Glue versions 1427
AWS Glue versions 1427
AWS Glue version support policy 1444
Migrating AWS Glue for Spark jobs to AWS Glue version 5.0 1446
Migrating AWS Glue for Spark jobs to AWS Glue version 4.0 1461
Upgrade analysis with AI 1475
Working with Spark jobs 1490
Job parameters 1490
Spark and PySpark jobs 1500
Worker types 1641
Streaming ETL jobs 1646
Record matching with FindMatches 1660
Migrate Spark programs 1693
Working with Ray jobs 1701
Getting started with AWS Glue for Ray 1701
Supported Ray runtime environments 1703
Accounting for workers in Ray jobs 1703
Ray job parameters 1704
Ray job metrics 1707
Configuring Python shell job properties 1708
Limitations 1709
Defining job properties for Python shell jobs 1709
Supported libraries for Python shell jobs 1711
Providing your own Python library 1713
Use AWS CloudFormation with Python shell jobs in AWS Glue 1717
Migrating from Python shell 3.6 to Python shell 3.9 1717
Migrate from AWS Glue Python shell jobs 1717
Monitoring 1719
AWS tags 1720
Automating with EventBridge 1725
Monitoring resources 1728
Logging using CloudTrail 1731
AWS Glue Streaming 1735
Use cases for streaming 1735
What are the benefits of using AWS Glue Streaming? 1736

- When to use AWS Glue Streaming? ... 1737
- Supported data sources ... 1738
- Supported data targets ... 1738
- Tutorial: Build your first streaming workload using AWS Glue Studio ... 1738
  - Prerequisites ... 1739
  - Consume streaming data from Amazon Kinesis ... 1739
- Tutorial: Build your first streaming workload using AWS Glue Studio notebooks ... 1750
  - Prerequisites ... 1751
  - Consume streaming data from Amazon Kinesis ... 1751
- Streaming concepts ... 1758
  - Anatomy of a AWS Glue streaming job ... 1758
- Streaming connections ... 1762
  - Kafka connections ... 1762
  - Kinesis connections ... 1769
- AWS Glue streaming autoscaling ... 1775
  - Enabling Auto Scaling in AWS Glue Studio ... 1776
  - Enabling Auto Scaling with the AWS CLI or SDK ... 1777
  - How it works ... 1777
- Maintenance windows ... 1779
  - Setting up a maintenance window ... 1779
  - Maintenance window behavior ... 1780
  - Job monitoring ... 1781
  - Data loss handling ... 1783
- Advanced AWS Glue streaming concepts ... 1784
  - Time considerations when processing streams ... 1784
  - Windowing ... 1785
  - Handling late data and watermarks ... 1790
- Monitoring AWS Glue streaming jobs ... 1792
  - Visualizing AWS Glue streaming metrics ... 1793
  - Using AWS Glue streaming metrics ... 1794
  - How to get the best performance ... 1799

Zero-ETL integrations ... 1801
- Zero-ETL capabilities in AWS Glue ... 1801
- Prerequisites ... 1802
  - Setting up source resources ... 1802
  - Setting up target resources ... 1804

    - Creating an Amazon Redshift data warehouse 1809
    - Setting up a VPC 1810
    - Setting up cross-account integration 1813
- Configuring a source 1814
    - Support for Special SAP Entities 1814
    - Amazon DynamoDB source 1814
    - Salesforce source 1816
    - Salesforce MCAE source 1817
    - Additional Salesforce Configuration 1817
    - SAP OData source 1817
    - ServiceNow source 1818
    - Zendesk source 1818
    - Zoho CRM source 1818
    - Facebook Ads source 1820
    - Instagram Ads source 1820
    - Unsupported Salesforce fields 1820
    - Unsupported ServiceNow fields 1823
- Configuring a target 1828
    - SageMaker Lakehouse with S3 storage 1829
    - S3 Tables target 1830
    - SageMaker Lakehouse with Amazon Redshift storage 1834
    - Amazon Redshift data warehouse target 1835
    - Configuring the integration target 1836
- Partition and schema unnesting 1837
    - Default unnesting & partitioning behavior 1838
    - Schema unnesting 1838
    - Data partitioning 1842
    - Limitations 1849
- Common tasks 1849
    - Creating an integration 1850
    - Modifying an integration 1851
    - Deleting an integration 1852
- Using APIs 1852
- Monitoring an integration 1853
    - Integration states 1853
    - Viewing logs 1853

    - Viewing metrics 1856
    - Event notifications 1856
  - Limitations 1857

AWS Glue Data Quality 1867

  - Benefits and key features 1867
  - How it works 1868
    - Data quality for the AWS Glue Data Catalog 1868
    - Data quality for AWS Glue ETL jobs 1869
    - Comparing AWS Glue Data Quality entry points 1869
  - Considerations 1871
  - Terminology 1871
  - Limits 1872
  - Release notes for AWS Glue Data Quality 1872
    - General availability: new features 1873
    - Nov 27, 2023 (Preview) 1873
    - Mar 12, 2024 1873
    - June 26, 2024 1873
    - August 7, 2024 1874
    - Nov 22, 2024 1874
    - Dec 6, 2024 1874
    - Jul 7, 2025 1874
  - Anomaly detection in AWS Glue Data Quality 1874
    - How it works 1875
    - Details of the Anomaly Detection algorithm 1880
  - IAM permissions for AWS Glue Data Quality 1881
    - IAM permissions 1881
    - IAM setup required for scheduling evaluation runs 1884
    - Example IAM policies 1885
  - Getting started with AWS Glue Data Quality for the Data Catalog 1890
    - Prerequisites 1890
    - Step-by-step example 1891
    - Generating rule recommendations 1891
    - Monitoring rule recommendations 1893
    - Editing recommended rulesets 1893
    - Creating a new ruleset 1894
    - Running a ruleset to evaluate data quality 1896

  - Viewing the data quality score and results 1897
  - Supported source types 1898
  - Related topics 1900
- Evaluating data quality with AWS Glue Studio 1900
  - Benefits 1900
  - Evaluating data quality for ETL jobs in AWS Glue Studio 1901
  - Data Quality rule builder 1911
  - Configuring anomaly detection in AWS Glue ETL jobs 1916
  - Viewing data quality scores and anomalies 1919
- Data Quality for ETL jobs in AWS Glue Studio notebooks 1922
  - Prerequisites 1923
  - Creating an ETL job in AWS Glue Studio 1923
- Data Quality Definition Language (DQDL) reference 1928
  - Syntax 1930
  - Rule type reference 1945
- Using APIs to measure and manage data quality 2001
  - Prerequisites 2001
  - Working with AWS Glue Data Quality recommendations 2001
  - Working with AWS Glue Data Quality rulesets 2004
  - Working with AWS Glue Data Quality runs 2007
  - Working with AWS Glue Data Quality results 2012
- Setting up alerts, deployments, and scheduling 2018
  - Setting up alerts and notifications in Amazon EventBridge integration 2018
  - Set up alerts and notifications in CloudWatch integration 2026
  - Querying data quality results 2027
  - Deploying data quality rules 2031
  - Scheduling data quality rules 2031
- Data encryption at rest for AWS Glue Data Quality 2032
  - AWS owned keys 2032
  - Customer managed keys 2032
  - Create a Customer Managed Key 2035
  - Create a security configuration 2037
  - AWS Glue Data Quality encryption context 2038
  - Monitoring your encryption keys for AWS Glue Data Quality 2039
  - Learn more 2043
- Troubleshooting AWS Glue Data Quality errors 2044

    - Error: missing module 2045
    - Error: insufficient permissions 2045
    - Error: rulesets not unique 2045
    - Error: tables with special characters 2045
    - Error: overflow with a large ruleset 2045
    - Error: rule status is failed 2046
    - AnalysisException: Unable to verify existence of default database 2046
    - Provided key map not suitable for given data frames 2046
    - java.lang.RuntimeException : Failed to fetch data. 2047
    - LAUNCH ERROR: Error downloading from S3 for bucket 2047
    - InvalidInputException (status: 400): DataQuality rules cannot be parsed 2047
    - Error: Eventbridge is not triggering Glue DQ jobs based on the schedule I setup 2048
    - CustomSQL errors 2048
    - Dynamic Rules 2049
    - Exception in User Class: org.apache.spark.sql.AnalysisException: org.apache.hadoop.hive.ql.metadata.HiveException 2051
    - UNCLASSIFIED_ERROR; IllegalArgumentException: Parsing Error: No rules or analyzers provided., no viable alternative at input 2051

Amazon Q data integration in AWS Glue 2052
- What is Amazon Q? 2052
- Amazon Q data integration in AWS Glue 2052
- Working with Amazon Q data integration 2053
- Best practices 2055
- Service improvement 2055
- Considerations 2056
- Setting up Amazon Q data integration 2056
    - Configuring IAM permissions 2056
- Supported code generation 2058
- Example interactions 2059
    - Amazon Q chat interactions 2059
    - AWS Glue Studio notebook interactions 2061
- Using context awareness 2064
    - Example: interactions 2064
    - Limitations 2068

Orchestration 2069
- Starting jobs and crawlers using triggers 2069

    - AWS Glue triggers 2069
    - Adding triggers 2072
    - Activating and deactivating triggers 2076
  - Performing complex ETL activities using blueprints and workflows 2077
    - Overview of workflows 2077
    - Creating and building out a workflow manually 2081
    - Starting a workflow with an EventBridge event 2085
    - Viewing the EventBridge events that started a workflow 2093
    - Running and monitoring a workflow 2094
    - Stopping a workflow run 2096
    - Repairing and resuming a workflow run 2097
    - Getting and setting workflow run properties 2103
    - Querying workflows using the AWS Glue API 2104
    - Blueprint and workflow restrictions 2109
    - Troubleshooting blueprint errors 2110
    - Permissions for blueprint personas and roles 2115
  - Developing blueprints 2119
    - Overview of blueprints 2120
    - Developing blueprints 2123
    - Registering a blueprint 2148
    - Viewing blueprints 2150
    - Updating a blueprint 2152
    - Creating a workflow from a blueprint 2154
    - Viewing blueprint runs 2155

AWS CloudFormation for AWS Glue 2157
  - Sample database 2159
  - Sample database, table, partitions 2160
  - Sample grok classifier 2164
  - Sample JSON classifier 2165
  - Sample XML classifier 2166
  - Sample Amazon S3 crawler 2167
  - Sample connection 2169
  - Sample JDBC crawler 2171
  - Sample job for Amazon S3 to Amazon S3 2173
  - Sample job for JDBC to Amazon S3 2175
  - Sample On-Demand trigger 2177

- Sample scheduled trigger 2178
- Sample conditional trigger 2179
- Sample machine learning transform 2180
- Sample data quality ruleset 2181
- Sample data quality ruleset with EventBridge scheduler 2183
- Sample development endpoint 2185

AWS Glue programming guide 2187

- Providing your own custom scripts 2187
- AWS Glue for Spark 2188
  - Tutorial: Writing a Spark script 2188
  - ETL in PySpark 2201
  - ETL in Scala 2446
  - Features and optimizations 2530
- AWS Glue for Ray 2788
  - Tutorial: Writing a Ray script 2789
  - Using Ray Core and Ray Data in AWS Glue for Ray 2794
  - Providing files and Python libraries 2796
  - Connecting to data 2801

Working with AWS SDKs 2803

AWS Glue API 2805

- Security 2833
  - — data types — 2833
  - DataCatalogEncryptionSettings 2833
  - EncryptionAtRest 2834
  - ConnectionPasswordEncryption 2834
  - EncryptionConfiguration 2835
  - S3Encryption 2836
  - CloudWatchEncryption 2836
  - JobBookmarksEncryption 2836
  - SecurityConfiguration 2837
  - GluePolicy 2837
  - DataQualityEncryption 2838
  - — operations — 2838
  - GetDataCatalogEncryptionSettings (get_data_catalog_encryption_settings) 2839
  - PutDataCatalogEncryptionSettings (put_data_catalog_encryption_settings) 2839
  - PutResourcePolicy (put_resource_policy) 2840

GetResourcePolicy (get_resource_policy) 2841
DeleteResourcePolicy (delete_resource_policy) 2842
CreateSecurityConfiguration (create_security_configuration) 2843
DeleteSecurityConfiguration (delete_security_configuration) 2844
GetSecurityConfiguration (get_security_configuration) 2845
GetSecurityConfigurations (get_security_configurations) 2845
GetResourcePolicies (get_resource_policies) 2846
Catalog objects 2847
Catalogs 2847
Databases 2862
Tables 2872
Partitions 2919
Connections 2945
User-defined Functions 2988
Importing an Athena catalog 2995
Table optimizer 2997
— data types — 2998
TableOptimizer 2998
TableOptimizerConfiguration 2999
TableOptimizerVpcConfiguration 3000
CompactionConfiguration 3000
IcebergCompactionConfiguration 3000
TableOptimizerRun 3001
BatchGetTableOptimizerEntry 3003
BatchTableOptimizer 3003
BatchGetTableOptimizerError 3004
RetentionConfiguration 3005
IcebergRetentionConfiguration 3005
OrphanFileDeletionConfiguration 3006
IcebergOrphanFileDeletionConfiguration 3006
CompactionMetrics 3006
RetentionMetrics 3007
OrphanFileDeletionMetrics 3007
IcebergCompactionMetrics 3007
IcebergRetentionMetrics 3007
IcebergOrphanFileDeletionMetrics 3008

- RunMetrics 3008
- — operations — 3009
- GetTableOptimizer (get_table_optimizer) 3009
- BatchGetTableOptimizer (batch_get_table_optimizer) 3010
- ListTableOptimizerRuns (list_table_optimizer_runs) 3011
- CreateTableOptimizer (create_table_optimizer) 3013
- DeleteTableOptimizer (delete_table_optimizer) 3014
- UpdateTableOptimizer (update_table_optimizer) 3015

Crawlers and classifiers 3016
- Classifiers 3016
- Crawlers 3030
- Column statistics 3058
- Scheduler 3074

Autogenerating ETL Scripts 3077
- — data types — 3077
- CodeGenNode 3077
- CodeGenNodeArg 3078
- CodeGenEdge 3078
- Location 3079
- CatalogEntry 3079
- MappingEntry 3079
- — operations — 3080
- CreateScript (create_script) 3080
- GetDataflowGraph (get_dataflow_graph) 3081
- GetMapping (get_mapping) 3082
- GetPlan (get_plan) 3083

Visual job API 3084
- — data types — 3084
- CodeGenConfigurationNode 3088
- JDBCConnectorOptions 3095
- StreamingDataPreviewOptions 3096
- AthenaConnectorSource 3097
- JDBCConnectorSource 3098
- SparkConnectorSource 3098
- CatalogSource 3099
- MySQLCatalogSource 3100

PostgreSQLCatalogSource 3100
OracleSQLCatalogSource 3101
MicrosoftSQLServerCatalogSource 3101
CatalogKinesisSource 3101
DirectKinesisSource 3102
KinesisStreamingSourceOptions 3103
CatalogKafkaSource 3106
DirectKafkaSource 3106
KafkaStreamingSourceOptions 3107
RedshiftSource 3109
AmazonRedshiftSource 3110
AmazonRedshiftNodeData 3110
AmazonRedshiftAdvancedOption 3113
Option 3113
S3CatalogSource 3113
S3SourceAdditionalOptions 3114
S3CsvSource 3114
DirectJDBCSource 3117
S3DirectSourceAdditionalOptions 3117
S3JsonSource 3118
S3ParquetSource 3120
S3DeltaSource 3121
S3CatalogDeltaSource 3122
CatalogDeltaSource 3122
S3HudiSource 3123
S3CatalogHudiSource 3124
S3ExcelSource 3124
CatalogHudiSource 3126
DynamoDBCatalogSource 3126
RelationalCatalogSource 3127
JDBCConnectorTarget 3127
SparkConnectorTarget 3128
BasicCatalogTarget 3129
MySQLCatalogTarget 3130
PostgreSQLCatalogTarget 3130
OracleSQLCatalogTarget 3131

MicrosoftSQLServerCatalogTarget 3131
RedshiftTarget 3132
AmazonRedshiftTarget 3132
UpsertRedshiftTargetOptions 3133
S3CatalogTarget 3133
S3GlueParquetTarget 3134
CatalogSchemaChangePolicy 3135
S3DirectTarget 3135
S3HudiCatalogTarget 3136
S3HudiDirectTarget 3137
S3DeltaCatalogTarget 3139
S3DeltaDirectTarget 3140
S3HyperDirectTarget 3141
S3IcebergDirectTarget 3142
DirectSchemaChangePolicy 3143
ApplyMapping 3143
Mapping 3144
SelectFields 3145
DropFields 3145
RenameField 3146
Spigot 3146
Join 3147
JoinColumn 3147
SplitFields 3148
SelectFromCollection 3148
FillMissingValues 3149
Filter 3149
FilterExpression 3150
FilterValue 3150
CustomCode 3151
SparkSQL 3151
SqlAlias 3152
DropNullFields 3152
NullCheckBoxList 3153
NullValueField 3153
Datatype 3154

Merge 3154
Union 3155
PIIDetection 3155
Aggregate 3157
DropDuplicates 3157
GovernedCatalogTarget 3158
GovernedCatalogSource 3158
AggregateOperation 3159
GlueSchema 3160
GlueStudioSchemaColumn 3160
GlueStudioColumn 3160
DynamicTransform 3162
TransformConfigParameter 3162
EvaluateDataQuality 3163
DQResultsPublishingOptions 3164
DQStopJobOnFailureOptions 3164
EvaluateDataQualityMultiFrame 3165
Recipe 3166
RecipeReference 3166
SnowflakeNodeData 3166
SnowflakeSource 3169
SnowflakeTarget 3169
ConnectorDataSource 3169
ConnectorDataTarget 3170
RecipeStep 3171
RecipeAction 3172
ConditionExpression 3172
S3CatalogIcebergSource 3173
CatalogIcebergSource 3173
S3IcebergCatalogTarget 3174
DynamoDBELTConnectorSource 3175
DDBELTConnectionOptions 3175
DDBELTCatalogAdditionalOptions 3176
Route 3177
GroupFilters 3177
AutoDataQuality 3178

Jobs 3178
  Jobs 3178
  Job runs 3204
  Triggers 3223
Integration APIs 3237
  — data types — 3237
  Integration 3238
  IntegrationConfig 3239
  IntegrationPartition 3240
  IntegrationError 3241
  IntegrationFilter 3242
  InboundIntegration 3242
  SourceProcessingProperties 3243
  TargetProcessingProperties 3243
  SourceTableConfig 3244
  TargetTableConfig 3244
  — operations — 3245
  CreateIntegration (create_integration) 3245
  ModifyIntegration (modify_integration) 3248
  DescribeIntegrations (describe_integrations) 3251
  DeleteIntegration (delete_integration) 3252
  DescribeInboundIntegrations (describe_inbound_integrations) 3254
  CreateIntegrationTableProperties (create_integration_table_properties) 3255
  UpdateIntegrationTableProperties (update_integration_table_properties) 3257
  GetIntegrationTableProperties (get_integration_table_properties) 3258
  DeleteIntegrationTableProperties (delete_integration_table_properties) 3259
  CreateIntegrationResourceProperty (create_integration_resource_property) 3260
  UpdateIntegrationResourceProperty (update_integration_resource_property) 3261
  GetIntegrationResourceProperty (get_integration_resource_property) 3262
  UntagResource (untag_resource) 3263
  ListTagsForResource (list_tags_for_resource) 3264
  — exceptions — 3264
  ResourceNotFoundException 3265
  InternalServerException 3265
  IntegrationAlreadyExistsFault 3265
  IntegrationConflictOperationFault 3265

IntegrationQuotaExceededFault 3266
KMSKeyNotAccessibleFault 3266
IntegrationNotFoundFault 3266
TargetResourceNotFound 3266
InvalidIntegrationStateFault 3267
Interactive sessions 3267
— data types — 3267
Session 3267
SessionCommand 3270
Statement 3270
StatementOutput 3271
StatementOutputData 3272
ConnectionsList 3272
— operations — 3272
CreateSession (create_session) 3273
StopSession (stop_session) 3276
DeleteSession (delete_session) 3277
GetSession (get_session) 3278
ListSessions (list_sessions) 3279
RunStatement (run_statement) 3280
CancelStatement (cancel_statement) 3281
GetStatement (get_statement) 3282
ListStatements (list_statements) 3282
GetGlueIdentityCenterConfiguration (get_glue_identity_center_configuration) 3284
UpdateGlueIdentityCenterConfiguration (update_glue_identity_center_configuration) ... 3285
CreateGlueIdentityCenterConfiguration (create_glue_identity_center_configuration) 3285
DeleteGlueIdentityCenterConfiguration (delete_glue_identity_center_configuration) 3286
DevEndpoints 3287
— data types — 3287
DevEndpoint 3287
DevEndpointCustomLibraries 3291
— operations — 3292
CreateDevEndpoint (create_dev_endpoint) 3292
UpdateDevEndpoint (update_dev_endpoint) 3298
DeleteDevEndpoint (delete_dev_endpoint) 3299
GetDevEndpoint (get_dev_endpoint) 3300

GetDevEndpoints (get_dev_endpoints) 3301
BatchGetDevEndpoints (batch_get_dev_endpoints) 3302
ListDevEndpoints (list_dev_endpoints) 3303
Schema registry 3304
— data types — 3304
RegistryId 3304
RegistryListItem 3305
MetadataInfo 3305
OtherMetadataValueListItem 3306
SchemaListItem 3306
SchemaVersionListItem 3307
MetadataKeyValuePair 3308
SchemaVersionErrorItem 3308
ErrorDetails 3308
SchemaVersionNumber 3309
SchemaId 3309
— operations — 3310
CreateRegistry (create_registry) 3310
CreateSchema (create_schema) 3312
GetSchema (get_schema) 3316
ListSchemaVersions (list_schema_versions) 3318
GetSchemaVersion (get_schema_version) 3319
GetSchemaVersionsDiff (get_schema_versions_diff) 3320
ListRegistries (list_registries) 3322
ListSchemas (list_schemas) 3322
RegisterSchemaVersion (register_schema_version) 3324
UpdateSchema (update_schema) 3325
CheckSchemaVersionValidity (check_schema_version_validity) 3327
UpdateRegistry (update_registry) 3327
GetSchemaByDefinition (get_schema_by_definition) 3328
GetRegistry (get_registry) 3330
PutSchemaVersionMetadata (put_schema_version_metadata) 3331
QuerySchemaVersionMetadata (query_schema_version_metadata) 3333
RemoveSchemaVersionMetadata (remove_schema_version_metadata) 3334
DeleteRegistry (delete_registry) 3336
DeleteSchema (delete_schema) 3337

DeleteSchemaVersions (delete_schema_versions) 3338
Workflows 3339
— data types — 3339
JobNodeDetails 3339
CrawlerNodeDetails 3340
TriggerNodeDetails 3340
Crawl 3340
Node 3341
Edge 3342
Workflow 3342
WorkflowGraph 3343
WorkflowRun 3344
WorkflowRunStatistics 3345
StartingEventBatchCondition 3346
Blueprint 3346
BlueprintDetails 3347
LastActiveDefinition 3348
BlueprintRun 3348
— operations — 3350
CreateWorkflow (create_workflow) 3351
UpdateWorkflow (update_workflow) 3352
DeleteWorkflow (delete_workflow) 3353
GetWorkflow (get_workflow) 3354
ListWorkflows (list_workflows) 3355
BatchGetWorkflows (batch_get_workflows) 3355
GetWorkflowRun (get_workflow_run) 3356
GetWorkflowRuns (get_workflow_runs) 3357
GetWorkflowRunProperties (get_workflow_run_properties) 3358
PutWorkflowRunProperties (put_workflow_run_properties) 3359
CreateBlueprint (create_blueprint) 3360
UpdateBlueprint (update_blueprint) 3361
DeleteBlueprint (delete_blueprint) 3362
ListBlueprints (list_blueprints) 3363
BatchGetBlueprints (batch_get_blueprints) 3364
StartBlueprintRun (start_blueprint_run) 3365
GetBlueprintRun (get_blueprint_run) 3365

GetBlueprintRuns (get_blueprint_runs) 3366
StartWorkflowRun (start_workflow_run) 3367
StopWorkflowRun (stop_workflow_run) 3368
ResumeWorkflowRun (resume_workflow_run) 3369
Usage profiles 3370
— data types — 3370
ProfileConfiguration 3370
ConfigurationObject 3371
UsageProfileDefinition 3372
— operations — 3372
CreateUsageProfile (create_usage_profile) 3372
GetUsageProfile (get_usage_profile) 3373
UpdateUsageProfile (update_usage_profile) 3375
DeleteUsageProfile (delete_usage_profile) 3375
ListUsageProfiles (list_usage_profiles) 3376
Machine learning 3377
— data types — 3377
TransformParameters 3378
EvaluationMetrics 3378
MLTransform 3378
FindMatchesParameters 3382
FindMatchesMetrics 3383
ConfusionMatrix 3384
GlueTable 3385
TaskRun 3386
TransformFilterCriteria 3387
TransformSortCriteria 3388
TaskRunFilterCriteria 3388
TaskRunSortCriteria 3389
TaskRunProperties 3389
FindMatchesTaskRunProperties 3390
ImportLabelsTaskRunProperties 3390
ExportLabelsTaskRunProperties 3390
LabelingSetGenerationTaskRunProperties 3391
SchemaColumn 3391
TransformEncryption 3391

MLUserDataEncryption 3392
ColumnImportance 3392
— operations — 3393
CreateMLTransform (create_ml_transform) 3393
UpdateMLTransform (update_ml_transform) 3397
DeleteMLTransform (delete_ml_transform) 3399
GetMLTransform (get_ml_transform) 3400
GetMLTransforms (get_ml_transforms) 3403
ListMLTransforms (list_ml_transforms) 3404
StartMLEvaluationTaskRun (start_ml_evaluation_task_run) 3405
StartMLLabelingSetGenerationTaskRun (start_ml_labeling_set_generation_task_run) 3406
GetMLTaskRun (get_ml_task_run) 3407
GetMLTaskRuns (get_ml_task_runs) 3409
CancelMLTaskRun (cancel_ml_task_run) 3410
StartExportLabelsTaskRun (start_export_labels_task_run) 3411
StartImportLabelsTaskRun (start_import_labels_task_run) 3412
Data Quality 3413
— data types — 3413
DataSource 3414
DataQualityRulesetListDetails 3415
DataQualityTargetTable 3415
DataQualityRulesetEvaluationRunDescription 3416
DataQualityRulesetEvaluationRunFilter 3417
DataQualityEvaluationRunAdditionalRunOptions 3417
DataQualityRuleRecommendationRunDescription 3417
DataQualityRuleRecommendationRunFilter 3418
DataQualityResult 3418
DataQualityAnalyzerResult 3420
DataQualityObservation 3421
MetricBasedObservation 3421
DataQualityMetricValues 3422
DataQualityRuleResult 3422
DataQualityResultDescription 3423
DataQualityResultFilterCriteria 3424
DataQualityRulesetFilterCriteria 3425
DataQualityAggregatedMetrics 3425

StatisticAnnotation 3426
TimestampedInclusionAnnotation 3427
AnnotationError 3427
DatapointInclusionAnnotation 3427
StatisticSummaryList 3428
StatisticSummary 3428
RunIdentifier 3429
StatisticModelResult 3430
DataQualityGlueTable 3430
— operations — 3431
StartDataQualityRulesetEvaluationRun (start_data_quality_ruleset_evaluation_run) 3433
CancelDataQualityRulesetEvaluationRun (cancel_data_quality_ruleset_evaluation_run) . 3434
GetDataQualityRulesetEvaluationRun (get_data_quality_ruleset_evaluation_run) 3435
ListDataQualityRulesetEvaluationRuns (list_data_quality_ruleset_evaluation_runs) 3437
StartDataQualityRuleRecommendationRun (start_data_quality_rule_recommendation_run) 3438
CancelDataQualityRuleRecommendationRun (cancel_data_quality_rule_recommendation_run) 3439
GetDataQualityRuleRecommendationRun (get_data_quality_rule_recommendation_run) 3440
ListDataQualityRuleRecommendationRuns (list_data_quality_rule_recommendation_runs) 3442
GetDataQualityResult (get_data_quality_result) 3443
BatchGetDataQualityResult (batch_get_data_quality_result) 3445
ListDataQualityResults (list_data_quality_results) 3445
CreateDataQualityRuleset (create_data_quality_ruleset) 3446
DeleteDataQualityRuleset (delete_data_quality_ruleset) 3448
GetDataQualityRuleset (get_data_quality_ruleset) 3449
ListDataQualityRulesets (list_data_quality_rulesets) 3450
UpdateDataQualityRuleset (update_data_quality_ruleset) 3451
ListDataQualityStatistics (list_data_quality_statistics) 3452
TimestampFilter 3453
CreateDataQualityRulesetRequest 3454
GetDataQualityRulesetResponse 3455
GetDataQualityResultResponse 3456
StartDataQualityRuleRecommendationRunRequest 3457

GetDataQualityRuleRecommendationRunResponse ... 3458
BatchPutDataQualityStatisticAnnotation (batch_put_data_quality_statistic_annotation) 3460
GetDataQualityModel (get_data_quality_model) ... 3461
GetDataQualityModelResult (get_data_quality_model_result) ... 3462
ListDataQualityStatisticAnnotations (list_data_quality_statistic_annotations) ... 3462
PutDataQualityProfileAnnotation (put_data_quality_profile_annotation) ... 3464
Sensitive Data ... 3464
— data types — ... 3464
CustomEntityType ... 3465
— operations — ... 3465
CreateCustomEntityType (create_custom_entity_type) ... 3465
DeleteCustomEntityType (delete_custom_entity_type) ... 3467
GetCustomEntityType (get_custom_entity_type) ... 3467
BatchGetCustomEntityTypes (batch_get_custom_entity_types) ... 3468
ListCustomEntityTypes (list_custom_entity_types) ... 3469
Tagging APIs ... 3470
— data types — ... 3470
Tag ... 3470
— operations — ... 3471
TagResource (tag_resource) ... 3471
UntagResource (untag_resource) ... 3472
GetTags (get_tags) ... 3472
Common data types ... 3473
Tag ... 3473
DecimalNumber ... 3474
ErrorDetail ... 3474
PropertyPredicate ... 3474
ResourceUri ... 3475
ColumnStatistics ... 3475
ColumnStatisticsError ... 3476
ColumnError ... 3476
ColumnStatisticsData ... 3476
BooleanColumnStatisticsData ... 3477
DateColumnStatisticsData ... 3478
DecimalColumnStatisticsData ... 3478
DoubleColumnStatisticsData ... 3479

- LongColumnStatisticsData 3479
- StringColumnStatisticsData 3480
- BinaryColumnStatisticsData 3480
- String patterns 3481

Exceptions 3483

- AccessDeniedException 3483
- AlreadyExistsException 3484
- ConcurrentModificationException 3484
- ConcurrentRunsExceededException 3484
- CrawlerNotRunningException 3484
- CrawlerRunningException 3485
- CrawlerStoppingException 3485
- EntityNotFoundException 3485
- FederationSourceException 3485
- FederationSourceRetryableException 3486
- GlueEncryptionException 3486
- IdempotentParameterMismatchException 3486
- IllegalWorkflowStateException 3487
- InternalServiceException 3487
- InvalidExecutionEngineException 3487
- InvalidInputException 3487
- InvalidStateException 3488
- InvalidTaskStatusTransitionException 3488
- JobDefinitionErrorException 3488
- JobRunInTerminalStateException 3488
- JobRunInvalidStateTransitionException 3489
- JobRunNotInTerminalStateException 3489
- LateRunnerException 3490
- NoScheduleException 3490
- OperationTimeoutException 3490
- ResourceNotReadyException 3490
- ResourceNumberLimitExceededException 3491
- SchedulerNotRunningException 3491
- SchedulerRunningException 3491
- SchedulerTransitioningException 3491
- UnrecognizedRunnerException 3492

    - ValidationException 3492
    - VersionMismatchException 3492
- **AWS Glue API code examples 3493**
  - Basics 3503
    - Hello AWS Glue 3504
    - Learn the basics 3513
    - Actions 3651
- **Security 3793**
  - Data protection 3794
    - Encrypting data at rest 3794
    - Encrypting data in transit 3812
    - FIPS compliance 3813
    - Key management 3813
    - AWS Glue dependency on other AWS services 3813
    - Development endpoints 3814
  - Identity and access management 3815
    - Audience 3816
    - Authenticating with identities 3816
    - Managing access using policies 3819
    - How AWS Glue works with IAM 3822
    - Configuring IAM permissions for AWS Glue 3829
    - AWS Glue access control policy examples 3862
    - Granting AWS managed policies 3889
    - Granting dynamically scoped policies 3897
    - Specifying AWS Glue Resource ARNs 3898
    - Granting cross-account access 3907
    - Troubleshooting 3915
  - AWS Lake Formation access control models 3917
    - Using AWS Glue with AWS Lake Formation for Full Table Access 3917
    - Lake Formation for FGAC 3929
  - Using Amazon S3 Access Grants with AWS Glue 3940
    - How AWS Glue works with S3 Access Grants 3941
    - S3 Access Grants considerations with AWS Glue 3941
    - Set up S3 Access Grants with AWS Glue 3942
  - Trusted Identity Propagation 3944
    - Overview 3945

    - Features and benefits 3945
    - Use cases 3945
    - How it works 3946
    - Integrations 3947
    - Getting started with trusted identity propagation in AWS Glue ETL 3948
    - Considerations and limitations for AWS Glue ETL Trusted Identity Propagation integration 3953
  - Logging and monitoring 3955
  - Compliance validation 3955
  - Resilience 3956
  - Infrastructure security 3957
    - Configuring interface VPC endpoints (AWS PrivateLink) for AWS Glue 3957
    - Configuring shared Amazon VPCs 3960
- **Troubleshooting AWS Glue 3961**
  - Gathering AWS Glue troubleshooting information 3961
  - Troubleshooting Spark errors 3962
    - Error: Resource unavailable 3963
    - Error: Could not find S3 endpoint or NAT gateway for subnetId in VPC 3963
    - Error: Inbound rule in security group required 3963
    - Error: Outbound rule in security group required 3964
    - Error: Job run failed because the role passed should be given assume role permissions for the AWS Glue service 3964
    - Error: DescribeVpcEndpoints action is unauthorized. unable to validate VPC ID vpc-id 3964
    - Error: DescribeRouteTables action is unauthorized. unable to validate subnet id: Subnet-id in VPC id: vpc-id 3964
    - Error: Failed to call ec2:DescribeSubnets 3965
    - Error: Failed to call ec2:DescribeSecurityGroups 3965
    - Error: Could not find subnet for AZ 3965
    - Error: Job run exception when writing to a JDBC target 3965
    - Error: Amazon S3: The operation is not valid for the object's storage class 3966
    - Error: Amazon S3 timeout 3966
    - Error: Amazon S3 access denied 3966
    - Error: Amazon S3 access key ID does not exist 3967
    - Error: Job run fails when accessing Amazon S3 with an `s3a://` URI 3967
    - Error: Amazon S3 service token expired 3969
    - Error: No private DNS for network interface found 3969

  - Error: Development endpoint provisioning failed ... 3969
  - Error: Notebook server CREATE_FAILED ... 3969
  - Error: Local notebook fails to start ... 3970
  - Error: Running crawler failed ... 3970
  - Error: Partitions were not updated ... 3970
  - Error: Job bookmark update failed due to version mismatch ... 3971
  - Error: A job is reprocessing data when job bookmarks are enabled ... 3971
  - Error: Failover behavior between VPCs in AWS Glue ... 3972
- Crawler errors when the crawler is using Lake Formation permissions ... 3973
  - Error: The S3 location: s3://examplepath is not registered ... 3973
  - Error: User/Role is not authorized to perform: lakeformation:GetDataAccess on resource 3974
  - Error: Insufficient Lake Formation permission(s) on (Database name: exampleDatabase, Table Name: exampleTable) ... 3974
  - Error: Insufficient Lake Formation permission(s) on s3://examplepath ... 3974
  - Frequently asked questions about crawler configuration using Lake Formation credentials ... 3975
- Troubleshooting Ray errors ... 3976
  - Inspecting Ray job logs ... 3976
  - Troubleshooting Ray job errors ... 3977
- AWS Glue machine learning exceptions ... 3978
  - CancelMLTaskRunActivity ... 3978
  - CreateMLTaskRunActivity ... 3979
  - DeleteMLTransformActivity ... 3980
  - GetMLTaskRunActivity ... 3980
  - GetMLTaskRunsActivity ... 3980
  - GetMLTransformActivity ... 3981
  - GetMLTransformsActivity ... 3981
  - GetSaveLocationForTransformArtifactActivity ... 3981
  - GetTaskRunArtifactActivity ... 3982
  - PublishMLTransformModelActivity ... 3982
  - PullLatestMLTransformModelActivity ... 3983
  - PutJobMetadataForMLTransformActivity ... 3983
  - StartExportLabelsTaskRunActivity ... 3984
  - StartImportLabelsTaskRunActivity ... 3984
  - StartMLEvaluationTaskRunActivity ... 3985
  - StartMLLabelingSetGenerationTaskRunActivity ... 3986

    - UpdateMLTransformActivity ... 3987
  - AWS Glue quotas ... 3988
- **Improving AWS Glue performance ... 3989**
  - Tuning strategies for your job type ... 3989
  - Improving Spark performance ... 3989
  - Optimizing reads with pushdown ... 3990
    - Predicate pushdown on files stored on Amazon S3 ... 3990
    - Pushdown when working with JDBC sources ... 3991
    - Notes and limitations for pushdown in AWS Glue ... 3994
  - Using auto scaling for AWS Glue ... 3994
    - Requirements ... 3995
    - Enabling Auto Scaling in AWS Glue Studio ... 1776
    - Enabling Auto Scaling with the AWS CLI or SDK ... 1777
    - Enabling Auto Scaling with Interactive sessions ... 3997
    - Tips and considerations ... 3998
    - Monitoring Auto Scaling with Amazon CloudWatch metrics ... 3998
    - Monitoring Auto Scaling with Amazon CloudWatch Logs ... 3999
    - Monitoring Auto Scaling with Spark UI ... 4000
    - Monitoring Auto Scaling job run DPU usage ... 4001
    - Limitations ... 4001
  - Workload partitioning with bounded execution ... 4001
    - Enabling workload partitioning ... 4001
    - Setting up an AWS Glue trigger to automatically run the job ... 4003
- **Known issues ... 4004**
  - Preventing cross-job data access ... 4004
- **Documentation history ... 4007**
  - Earlier updates ... 4063
- **AWS Glossary ... 4065**

What is AWS Glue?

AWS Glue is a serverless data integration service that makes it easy for analytics users to discover, prepare, move, and integrate data from multiple sources. You can use it for analytics, machine learning, and application development. It also includes additional productivity and data ops tooling for authoring, running jobs, and implementing business workflows.

With AWS Glue, you can discover and connect to more than 70 diverse data sources and manage your data in a centralized data catalog. You can visually create, run, and monitor extract, transform, and load (ETL) pipelines to load data into your data lakes. Also, you can immediately search and query cataloged data using Amazon Athena, Amazon EMR, and Amazon Redshift Spectrum.

AWS Glue consolidates major data integration capabilities into a single service. These include data discovery, modern ETL, cleansing, transforming, and centralized cataloging. It's also serverless, which means there's no infrastructure to manage. With flexible support for all workloads like ETL, ELT, and streaming in one service, AWS Glue supports users across various workloads and types of users.

Also, AWS Glue makes it easy to integrate data across your architecture. It integrates with AWS analytics services and Amazon S3 data lakes. AWS Glue has integration interfaces and job-authoring tools that are easy to use for all users, from developers to business users, with tailored solutions for varied technical skill sets.

With the ability to scale on demand, AWS Glue helps you focus on high-value activities that maximize the value of your data. It scales for any data size, and supports all data types and schema variances. To increase agility and optimize costs, AWS Glue provides built-in high availability and pay-as-you-go billing.

For pricing information, see AWS Glue pricing.

AWS Glue Studio

AWS Glue Studio is a graphical interface that makes it easy to create, run, and monitor data integration jobs in AWS Glue. You can visually compose data transformation workflows and seamlessly run them on the Apache Spark–based serverless ETL engine in AWS Glue.

With AWS Glue Studio, you can create and manage jobs that gather, transform, and clean data. You can also use AWS Glue Studio to troubleshoot and edit job scripts.

Topics

- AWS Glue features
- Learning about innovations in AWS Glue
- Getting started with AWS Glue
- Accessing AWS Glue
- Related services

AWS Glue features

AWS Glue features fall into three major categories:

- Discover and organize data
- Transform, prepare, and clean data for analysis
- Build and monitor data pipelines

Discover and organize data

- **Unify and search across multiple data stores** – Store, index, and search across multiple data sources and sinks by cataloging all your data in AWS.
- **Automatically discover data** – Use AWS Glue crawlers to automatically infer schema information and integrate it into your AWS Glue Data Catalog.
- **Manage schemas and permissions** – Validate and control access to your databases and tables.
- **Connect to a wide variety of data sources** – Tap into multiple data sources, both on premises and on AWS, using AWS Glue connections to build your data lake.

Transform, prepare, and clean data for analysis

- **Visually transform data with a job canvas interface** – Define your ETL process in the visual job editor and automatically generate the code to extract, transform, and load your data.
- **Build complex ETL pipelines with simple job scheduling** – Invoke AWS Glue jobs on a schedule, on demand, or based on an event.
- **Clean and transform streaming data in transit** – Enable continuous data consumption, and clean and transform it in transit. This makes it available for analysis in seconds in your target data store.

- **Deduplicate and cleanse data with built-in machine learning –** Clean and prepare your data for analysis without becoming a machine learning expert by using the `FindMatches` feature. This feature deduplicates and finds records that are imperfect matches for each other.
- **Built-in job notebooks –** AWS Glue job notebooks provide serverless notebooks with minimal setup in AWS Glue so you can get started quickly.
- **Edit, debug, and test ETL code –** With AWS Glue interactive sessions, you can interactively explore and prepare data. You can explore, experiment on, and process data interactively using the IDE or notebook of your choice.
- **Define, detect, and remediate sensitive data –** AWS Glue sensitive data detection lets you define, identify, and process sensitive data in your data pipeline and in your data lake.

Build and monitor data pipelines

- **Automatically scale based on workload –** Dynamically scale resources up and down based on workload. This assigns workers to jobs only when needed.
- **Automate jobs with event-based triggers –** Start crawlers or AWS Glue jobs with event-based triggers, and design a chain of dependent jobs and crawlers.
- **Run and monitor jobs –** Run AWS Glue jobs with your choice of engine, Spark or Ray. Monitor them with automated monitoring tools, AWS Glue job run insights, and AWS CloudTrail. Improve your monitoring of Spark-backed jobs with the Apache Spark UI.
- **Define workflows for ETL and integration activities –** Define workflows for ETL and integration activities for multiple crawlers, jobs, and triggers.

Learning about innovations in AWS Glue

Learn about the latest innovations in AWS Glue and hear how customers use AWS Glue to enable self-service data preparation across their organization.

Learn about how customers scale AWS Glue beyond the traditional setup and how they configure AWS Glue for job monitoring and performance.

Getting started with AWS Glue

We recommend that you start with the following sections:

- Overview of using AWS Glue

- AWS Glue concepts
- Setting up IAM permissions for AWS Glue
- Getting started with the AWS Glue Data Catalog
- Authoring jobs in AWS Glue
- Getting started with AWS Glue interactive sessions
- Orchestration in AWS Glue

Accessing AWS Glue

You can create, view, and manage your AWS Glue jobs using the following interfaces:

- **AWS Glue console** – Provides a web interface for you to create, view, and manage your AWS Glue jobs. To access the console, see AWS Glue.
- **AWS Glue Studio** – Provides a graphical interface for you to create and edit your AWS Glue jobs visually. For more information, see Building visual ETL jobs.
- **AWS Glue section of the AWS CLI Reference** – Provides AWS CLI commands that you can use with AWS Glue. For more information, see AWS CLI Reference for AWS Glue.
- **AWS Glue API** – Provides a complete API reference for developers. For more information, see AWS Glue API.

Related services

Users of AWS Glue also use:

- **AWS Lake Formation** – A service that is an authorization layer that provides fine-grained access control to resources in the AWS Glue Data Catalog.
- **AWS Glue DataBrew** – A visual data preparation tool that you can use to clean and normalize data without writing any code.

AWS Glue: How it works

AWS Glue uses other AWS services to orchestrate your ETL (extract, transform, and load) jobs to build data warehouses and data lakes and generate output streams. AWS Glue calls API operations to transform your data, create runtime logs, store your job logic, and create notifications to help you monitor your job runs. The AWS Glue console connects these services into a managed application, so you can focus on creating and monitoring your ETL work. The console performs administrative and job development operations on your behalf. You supply credentials and other properties to AWS Glue to access your data sources and write to your data targets.

AWS Glue takes care of provisioning and managing the resources that are required to run your workload. You don't need to create the infrastructure for an ETL tool because AWS Glue does it for you. When resources are required, to reduce startup time, AWS Glue uses an instance from its warm pool of instances to run your workload.

With AWS Glue, you create jobs using table definitions in your Data Catalog. Jobs consist of scripts that contain the instructions that execute the desired data transformation tasks. You use triggers to initiate jobs either on a schedule or as a result of a specified event. You determine where your target data resides and which source data populates your target. Based on your inputs, AWS Glue transforms your data from the source to the target format. Alternatively, you can also provide custom scripts in the AWS Glue console or API to process your data according to your specific requirements.

Data sources and destinations

AWS Glue for Spark allows you to read and write data from multiple systems and databases including:

- Amazon S3
- Amazon DynamoDB
- Amazon Redshift
- Amazon Relational Database Service (Amazon RDS)
- Third-party JDBC-accessible databases
- MongoDB and Amazon DocumentDB (with MongoDB compatibility)
- Other marketplace connectors and Apache Spark plugins

Data streams

AWS Glue for Spark can stream data from the following systems:

- Amazon Kinesis Data Streams
- Apache Kafka

AWS Glue is available in several AWS Regions. For more information, see AWS Regions and Endpoints in the Amazon Web Services General Reference.

Topics

- Serverless ETL jobs run in isolation
- AWS Glue concepts
- AWS Glue components
- AWS Glue for Spark and AWS Glue for Ray
- Converting semi-structured schemas to relational schemas with AWS Glue
- AWS Glue type systems

Serverless ETL jobs run in isolation

AWS Glue runs your ETL jobs in a serverless environment with your choice of engine, Spark or Ray. AWS Glue runs these jobs on virtual resources that it provisions and manages in its own service account.

AWS Glue is designed to do the following:

- Segregate customer data.
- Protect customer data in transit and at rest.
- Access customer data only as needed in response to customer requests, using temporary, scoped-down credentials, or with a customer's consent to IAM roles in their account.

During provisioning of an ETL job, you provide input data sources and output data targets in your virtual private cloud (VPC). In addition, you provide the IAM role, VPC ID, subnet ID, and security group that are needed to access data sources and targets. For each tuple (customer account ID, IAM role, subnet ID, and security group), AWS Glue creates a new environment that is isolated at the network and management level from all other environments inside your AWS Glue service account.

You create and configure AWS Glue resources, such as Data Catalogs, Jobs, and Crawlers within your AWS account. These resources are then associated with the IAM role and network settings (subnet and security group) you specify during the creation process.

AWS Glue creates elastic network interfaces in your subnet using private IP addresses. Jobs use these elastic network interfaces to access your data sources and data targets. Traffic in, out, and within the job run environment is governed by your VPC and networking policies with one exception: Calls made to AWS Glue libraries can proxy traffic to AWS Glue API operations through the AWS Glue VPC. All AWS Glue API calls are logged; thus, data owners can audit API access by enabling AWS CloudTrail, which delivers audit logs to your account.

AWS Glue managed environments that run your ETL jobs are protected with the same security practices followed by other AWS services. For an overview of the practices and shared security responsibilities, see the Introduction to AWS Security Processes whitepaper.

AWS Glue concepts

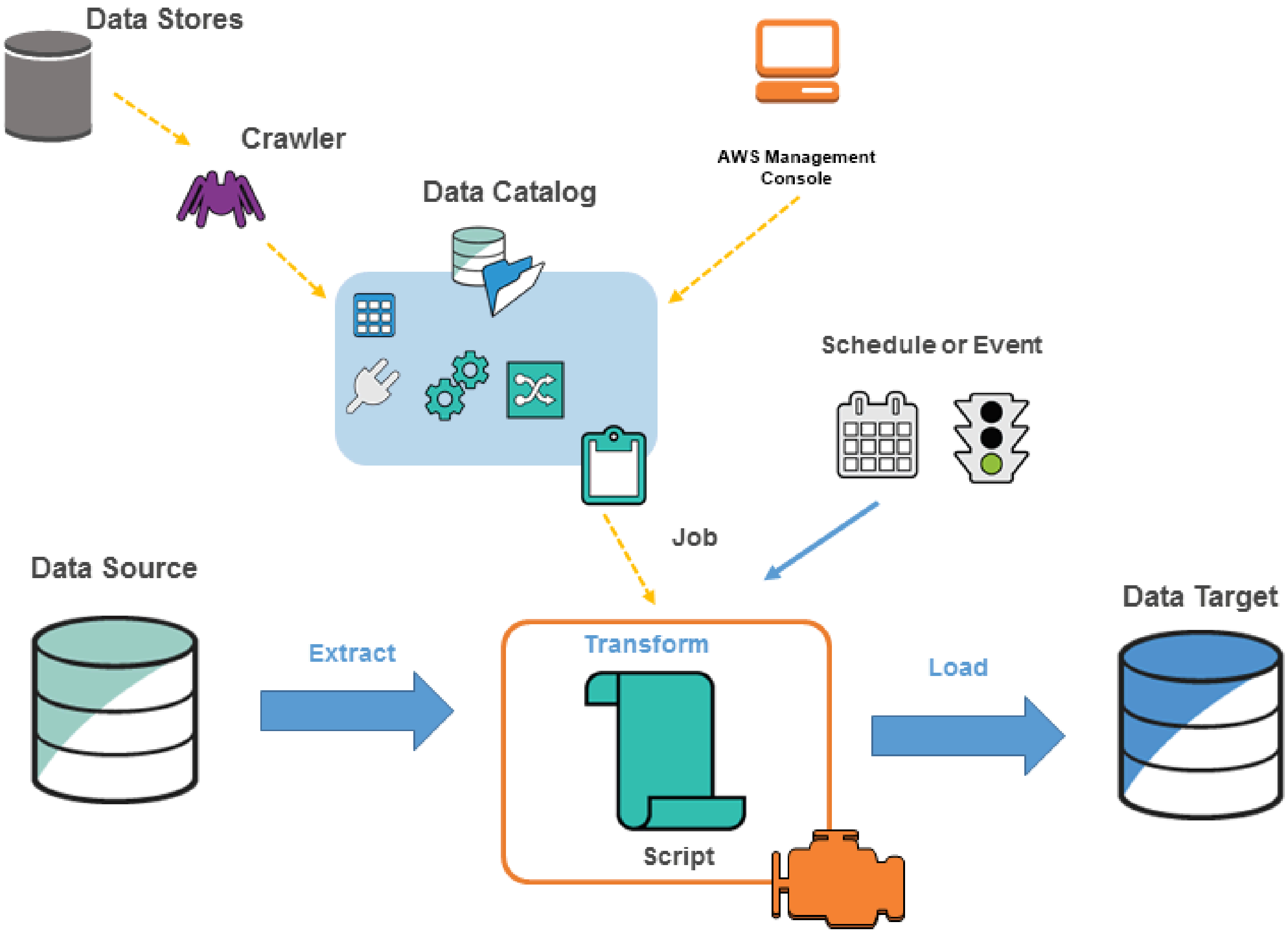
AWS Glue is a fully managed ETL (extract, transform, load) service that allows you to easily move data between different data sources and targets. The key components are:

- **Data Catalog**: A metadata store containing table definitions, job definitions, and other control information for your ETL workflows.
- **Crawlers**: Programs that connect to data sources, infer data schemas, and create metadata table definitions in the Data Catalog.
- **ETL Jobs**: The business logic to extract data from sources, transform it using Apache Spark scripts, and load it into targets.
- **Triggers**: Mechanisms to initiate job runs based on schedules or events.

The typical workflow involves:

1. Define data sources and targets in the Data Catalog.
2. Use Crawlers to populate the Data Catalog with table metadata from data sources.
3. Define ETL jobs with transformation scripts to move and process data.
4. Run jobs on-demand or based on triggers.
5. Monitor job performance using dashboards.

The following diagram shows the architecture of an AWS Glue environment.


Data Stores
Crawler
Data Catalog
AWS Management Console
Schedule or Event
Job
Data Source
Extract
Transform
Script
Load
Data Target

You define *jobs* in AWS Glue to accomplish the work that's required to extract, transform, and load (ETL) data from a data source to a data target. You typically perform the following actions:

- For data store sources, you define a *crawler* to populate your AWS Glue Data Catalog with metadata table definitions. You point your crawler at a data store, and the crawler creates table definitions in the Data Catalog. For streaming sources, you manually define Data Catalog tables and specify data stream properties.

  In addition to table definitions, the AWS Glue Data Catalog contains other metadata that is required to define ETL jobs. You use this metadata when you define a job to transform your data.

- AWS Glue can generate a script to transform your data. Or, you can provide the script in the AWS Glue console or API.

- You can run your job on demand, or you can set it up to start when a specified *trigger* occurs. The trigger can be a time-based schedule or an event.

  When your job runs, a script extracts data from your data source, transforms the data, and loads it to your data target. The script runs in an Apache Spark environment in AWS Glue.

> ⚠ **Important**
>
> Tables and databases in AWS Glue are objects in the AWS Glue Data Catalog. They contain metadata; they don't contain data from a data store.

Text-based data, such as CSVs, must be encoded in UTF-8 for AWS Glue to process it successfully. For more information, see UTF-8 in Wikipedia.

AWS Glue terminology

AWS Glue relies on the interaction of several components to create and manage your extract, transform, and load (ETL) workflow.

AWS Glue Data Catalog

The persistent metadata store in AWS Glue. It contains table definitions, job definitions, and other control information to manage your AWS Glue environment. Each AWS account has one AWS Glue Data Catalog per region.

Classifier

Determines the schema of your data. AWS Glue provides classifiers for common file types, such as CSV, JSON, AVRO, XML, and others. It also provides classifiers for common relational database management systems using a JDBC connection. You can write your own classifier by using a grok pattern or by specifying a row tag in an XML document.

Connection

A Data Catalog object that contains the properties that are required to connect to a particular data store.

Crawler

A program that connects to a data store (source or target), progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in the AWS Glue Data Catalog.

Database

A set of associated Data Catalog table definitions organized into a logical group.

Data store, data source, data target

A *data store* is a repository for persistently storing your data. Examples include Amazon S3 buckets and relational databases. A *data source* is a data store that is used as input to a process or transform. A *data target* is a data store that a process or transform writes to.

Development endpoint

An environment that you can use to develop and test your AWS Glue ETL scripts.

Dynamic Frame

A distributed table that supports nested data such as structures and arrays. Each record is self-describing, designed for schema flexibility with semi-structured data. Each record contains both data and the schema that describes that data. You can use both dynamic frames and Apache Spark DataFrames in your ETL scripts, and convert between them. Dynamic frames provide a set of advanced transformations for data cleaning and ETL.

Job

The business logic that is required to perform ETL work. It is composed of a transformation script, data sources, and data targets. Job runs are initiated by triggers that can be scheduled or triggered by events.

Job performance dashboard

AWS Glue provides a comprehensive run dashboard for your ETL jobs. The dashboard displays information about job runs from a specific time frame.

Notebook interface

An enhanced notebook experience with one-click setup for easy job authoring and data exploration. The notebook and connections are configured automatically for you. You can use the notebook interface based on Jupyter Notebook to interactively develop, debug, and deploy scripts and workflows using AWS Glue serverless Apache Spark ETL infrastructure. You can also perform ad-hoc queries, data analysis, and visualization (for example, tables and graphs) in the notebook environment.

Script

Code that extracts data from sources, transforms it, and loads it into targets. AWS Glue generates PySpark or Scala scripts.

Table

The metadata definition that represents your data. Whether your data is in an Amazon Simple Storage Service (Amazon S3) file, an Amazon Relational Database Service (Amazon RDS) table, or another set of data, a table defines the schema of your data. A table in the AWS Glue Data Catalog consists of the names of columns, data type definitions, partition information, and other metadata about a base dataset. The schema of your data is represented in your AWS Glue table definition. The actual data remains in its original data store, whether it be in a file or a relational database table. AWS Glue catalogs your files and relational database tables in the AWS Glue Data Catalog. They are used as sources and targets when you create an ETL job.

Transform

The code logic that is used to manipulate your data into a different format.

Trigger

Initiates an ETL job. Triggers can be defined based on a scheduled time or an event.

Visual job editor

The visual job editor is a graphical interface that makes it easy to create, run, and monitor extract, transform, and load (ETL) jobs in AWS Glue. You can visually compose data transformation workflows, seamlessly run them on AWS Glue's Apache Spark-based serverless ETL engine, and inspect the schema and data results in each step of the job.

Worker

With AWS Glue, you only pay for the time your ETL job takes to run. There are no resources to manage, no upfront costs, and you are not charged for startup or shutdown time. You are charged an hourly rate based on the number of **Data Processing Units** (or DPUs) used to run your ETL job. A single Data Processing Unit (DPU) is also referred to as a *worker*. AWS Glue comes with multiple worker types to help you select the configuration that meets your job latency and cost requirements. Workers come in Standard, G.1X, G.2X, G.4X, G.8X, G.12X, G.16X, G.025X, and memory-optimized R.1X, R.2X, R.4X, R.8X configurations.

AWS Glue components

AWS Glue provides a console and API operations to set up and manage your extract, transform, and load (ETL) workload. You can use API operations through several language-specific SDKs and the AWS Command Line Interface (AWS CLI). For information about using the AWS CLI, see AWS CLI Command Reference.

AWS Glue uses the AWS Glue Data Catalog to store metadata about data sources, transforms, and targets. The Data Catalog is a drop-in replacement for the Apache Hive Metastore. The AWS Glue Jobs system provides a managed infrastructure for defining, scheduling, and running ETL operations on your data. For more information about the AWS Glue API, see AWS Glue API.

AWS Glue console

You use the AWS Glue console to define and orchestrate your ETL workflow. The console calls several API operations in the AWS Glue Data Catalog and AWS Glue Jobs system to perform the following tasks:

- Define AWS Glue objects such as jobs, tables, crawlers, and connections.
- Schedule when crawlers run.
- Define events or schedules for job triggers.
- Search and filter lists of AWS Glue objects.
- Edit transformation scripts.

AWS Glue Data Catalog

The AWS Glue Data Catalog is your persistent technical metadata store in the AWS Cloud.

Each AWS account has one AWS Glue Data Catalog per AWS Region. Each Data Catalog is a highly scalable collection of tables organized into databases. A table is metadata representation of a collection of structured or semi-structured data stored in sources such as Amazon RDS, Apache Hadoop Distributed File System, Amazon OpenSearch Service, and others. The AWS Glue Data Catalog provides a uniform repository where disparate systems can store and find metadata to keep track of data in data silos. You can then use the metadata to query and transform that data in a consistent manner across a wide variety of applications.

You use the Data Catalog together with AWS Identity and Access Management policies and Lake Formation to control access to the tables and databases. By doing this, you can allow different groups in your enterprise to safely publish data to the wider organization while protecting sensitive information in a highly granular fashion.

The Data Catalog, along with CloudTrail and Lake Formation, also provides you with comprehensive audit and governance capabilities, with schema change tracking and data access controls. This helps ensure that data is not inappropriately modified or inadvertently shared.

For information about securing and auditing the AWS Glue Data Catalog, see:

- **AWS Lake Formation** – For more information, see What Is AWS Lake Formation? in the *AWS Lake Formation Developer Guide*.
- **CloudTrail** – For more information, see What Is CloudTrail? in the *AWS CloudTrail User Guide*.

The following are other AWS services and open-source projects that use the AWS Glue Data Catalog:

- **Amazon Athena** – For more information, see Understanding Tables, Databases, and the Data Catalog in the *Amazon Athena User Guide*.
- **Amazon Redshift Spectrum** – For more information, see Using Amazon Redshift Spectrum to Query External Data in the *Amazon Redshift Database Developer Guide*.
- **Amazon EMR** – For more information, see Use Resource-Based Policies for Amazon EMR Access to AWS Glue Data Catalog in the *Amazon EMR Management Guide*.
- **AWS Glue Data Catalog client for Apache Hive metastore** – For more information about this GitHub project, see AWS Glue Data Catalog Client for Apache Hive Metastore.

AWS Glue crawlers and classifiers

AWS Glue also lets you set up crawlers that can scan data in all kinds of repositories, classify it, extract schema information from it, and store the metadata automatically in the AWS Glue Data Catalog. The AWS Glue Data Catalog can then be used to guide ETL operations.

For information about how to set up crawlers and classifiers, see Using crawlers to populate the Data Catalog . For information about how to program crawlers and classifiers using the AWS Glue API, see Crawlers and classifiers API.

AWS Glue ETL operations

Using the metadata in the Data Catalog, AWS Glue can automatically generate Scala or PySpark (the Python API for Apache Spark) scripts with AWS Glue extensions that you can use and modify to perform various ETL operations. For example, you can extract, clean, and transform raw data, and then store the result in a different repository, where it can be queried and analyzed. Such a script might convert a CSV file into a relational form and save it in Amazon Redshift.

For more information about how to use AWS Glue ETL capabilities, see Programming Spark scripts.

Streaming ETL in AWS Glue

AWS Glue enables you to perform ETL operations on streaming data using continuously-running jobs. AWS Glue streaming ETL is built on the Apache Spark Structured Streaming engine, and can ingest streams from Amazon Kinesis Data Streams, Apache Kafka, and Amazon Managed Streaming for Apache Kafka (Amazon MSK). Streaming ETL can clean and transform streaming data and load it into Amazon S3 or JDBC data stores. Use Streaming ETL in AWS Glue to process event data like IoT streams, clickstreams, and network logs.

If you know the schema of the streaming data source, you can specify it in a Data Catalog table. If not, you can enable schema detection in the streaming ETL job. The job then automatically determines the schema from the incoming data.

The streaming ETL job can use both AWS Glue built-in transforms and transforms that are native to Apache Spark Structured Streaming. For more information, see Operations on streaming DataFrames/Datasets on the Apache Spark website.

For more information, see the section called “Streaming ETL jobs”.

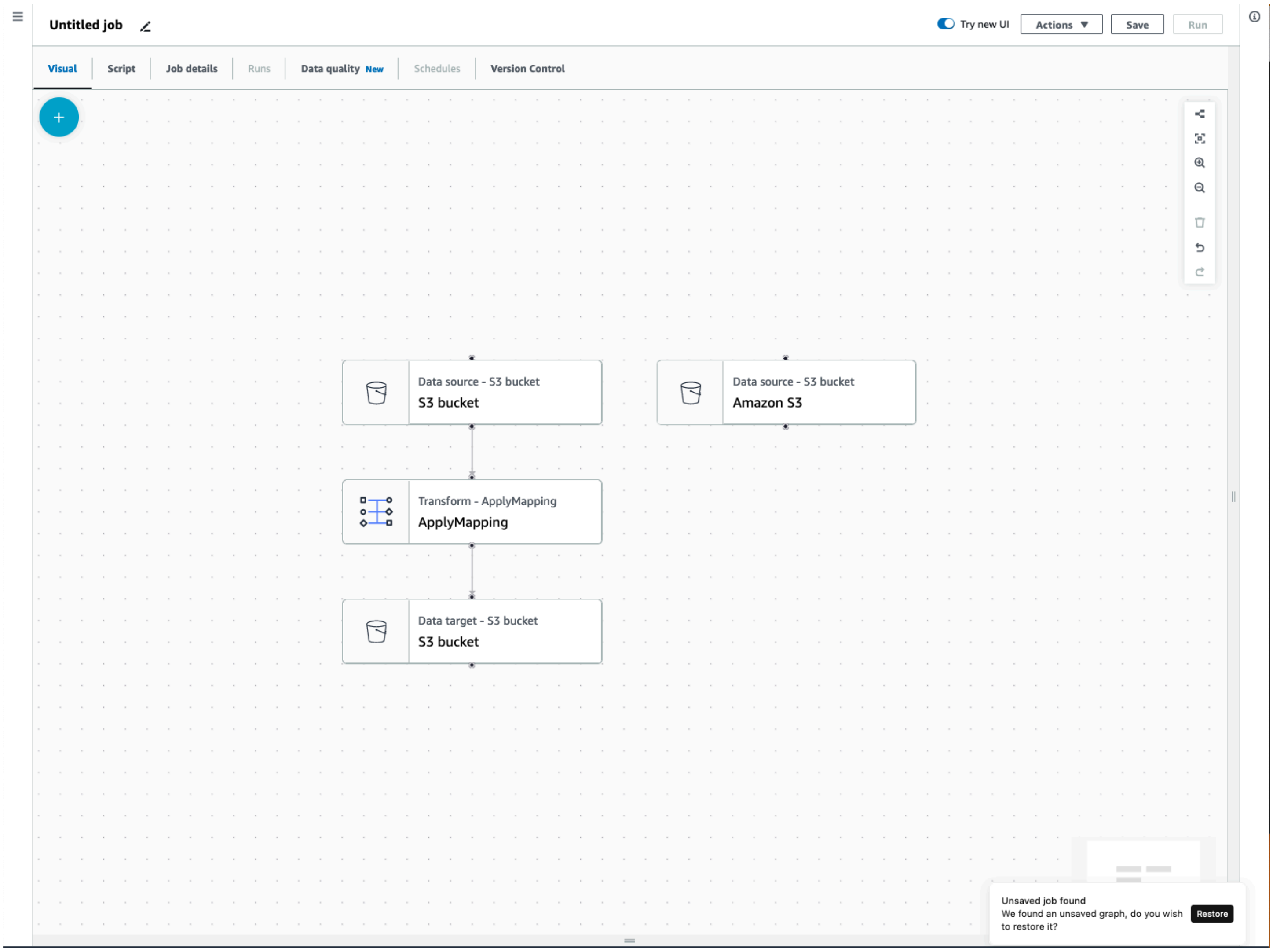
The AWS Glue jobs system

The AWS Glue Jobs system provides managed infrastructure to orchestrate your ETL workflow. You can create jobs in AWS Glue that automate the scripts you use to extract, transform, and transfer data to different locations. Jobs can be scheduled and chained, or they can be triggered by events such as the arrival of new data.

For more information about using the AWS Glue Jobs system, see Monitoring AWS Glue. For information about programming using the AWS Glue Jobs system API, see Jobs API.

Visual ETL components

AWS Glue allows you to create ETL jobs through a visual canvas that you can manipulate.


Untitled job
Try new UI
Actions
Save
Run
Visual
Script
Job details
Runs
Data quality New
Schedules
Version Control
Data source - S3 bucket
S3 bucket
Data source - S3 bucket
Amazon S3
Transform - ApplyMapping
ApplyMapping
Data target - S3 bucket
S3 bucket
Restore

Unsaved job found
We found an unsaved graph, do you wish to restore it?

ETL job menu

Menu options at the top of the canvas allow you to access the various views and configuration details about your job.

- **Visual –** The Visual job editor canvas. This is where you can add nodes to create a job.
- **Script –** The script representation of your ETL job. AWS Glue generates the script based on the visual representation of your job. You can also edit your script or download it.

> **Note**
>
> If you choose to edit the script, the job authoring experience is permanently converted to a script-only mode. Afterwards, you cannot use the visual editor to edit the job anymore. You should add all the job sources, transforms, and targets, and make all the changes you require with the visual editor before choosing to edit the script.

- **Job details –** The Job details tab allows you to configure your job by setting job properties. There are basic properties, such as name and description of your job, IAM role, job type, AWS Glue version, language, worker type, number of workers, job bookmark, flex execution, number of retires, and job timeout, and there are advanced properties, such as connections, libraries, job parameters, and tags.
- **Runs –** After your job runs, this tab can be accessed to view your past job runs.
- **Data quality –** Data quality evaluates and monitors the quality of your data assets. You can learn more about how to use data quality on this tab and add a data quality transform to your job.
- **Schedules –** Jobs that you've scheduled appear in this tab. If there are no schedules attached to this job, then this tab is not accessible.
- **Version control –** You can use Git with your job by configuring your job to a Git repository.

Visual ETL panels

When you work in the canvas, several panels are available to help you configure your nodes, or help you to preview your data and view the output schema.

- **Properties –** The Properties panel appears when you choose a node on your canvas.
- **Data preview –** The Data preview panel provides a preview of the data output so that you can make decisions before you run your job and examine your output.

- **Output schema –** The Output schema tab allows you to view and edit the schema of your transform nodes.

Resizing panels

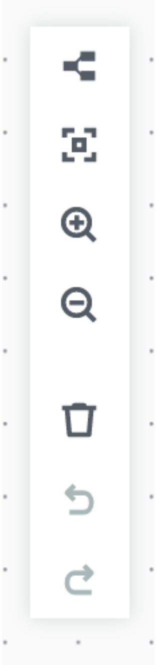
You can resize the Properties panel on the right-hand side of the screen and the bottom panel which contains the Data preview and Output schema tabs by clicking on the edge of the panel and dragging it left and right or up and down.

- **Properties panel –** Resize the properties panel by clicking and dragging the edge of the canvas on the right side of the screen and drag it left to expand its width. By default, the panel is collapsed and when a node is selected, the properties panel opens up to its default size.
- **Data preview and Output schema panel –** Resize the bottom panel by clicking and dragging the bottom edge of the canvas at the bottom of the screen and drag it up to expand its height. By default, the panel is collapsed and when a node is selected, the bottom panel opens up to its default size.

Job canvas

You can add, remove, and move/reorder nodes directly on the Visual ETL canvas. Think of it as your workspace to create a fully functional ETL job that starts with a data source and can end with a data target.

When you work with nodes on the canvas, you have a toolbar that can help you zoom in and out, remove nodes, make or edit connections between nodes, change the job flow orientation, and undo or redo an action.



The floating toolbar is anchored to the upper right-hand size of the canvas and contains several images that perform actions:

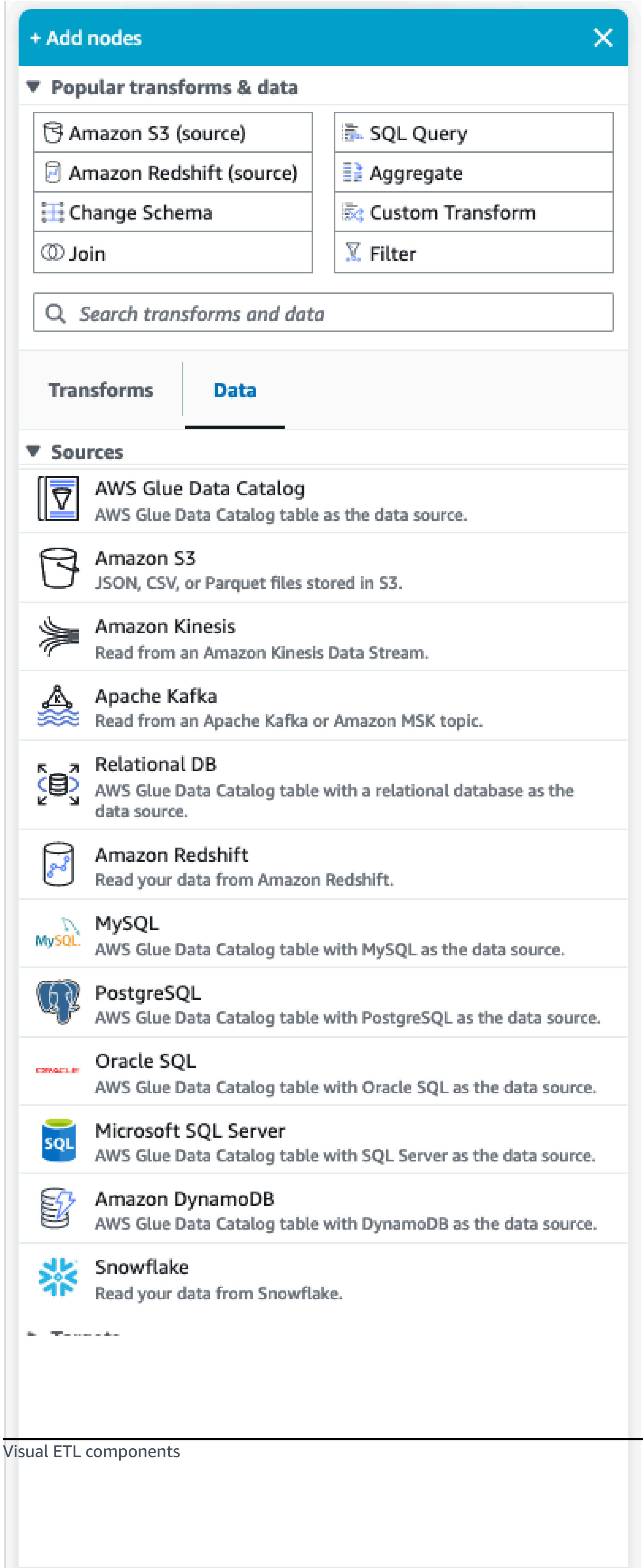
- **Layout icon** – The first icon in the toolbar is the layout icon. By default, the direction of visual jobs is top to bottom.It rearranges the direction of your visual job by arranging the nodes horizontally from left to right. Clicking the layout icon again changes the direction back to top to bottom.
- **Recenter icon** – The recenter icon changes the canvas view by centering it. You can use this with large jobs to get back to the center position.
- **Zoom in icon –** The zoom in icon enlarges the size of the nodes on the canvas.
- **Zoom out icon –** The zoom out icon decreases the size of the nodes on the canvas.
- **Trash icon –** The trash icon removes a node from the visual job. You must select a node first.
- **Undo icon –** The undo icon reverses the last action taken on the visual job.
- **Redo icon –** The redo icon repeats the last action taken on the visual job.

Using the mini-map

Resource panel

The resource panel contains all of the data sources, transform actions, and connections available to you. Open the resource panel on the canvas by clicking the "+" icon. This will open the resource panel.

To close the resource panel, click the **X** in the upper-right hand corner of the resource panel. This will hide the panel until you're ready to open it again.


+ Add nodes
Popular transforms & data
Amazon S3 (source)
SQL Query
Amazon Redshift (source)
Aggregate
Change Schema
Custom Transform
Join
Filter
Search transforms and data
Transforms
Data
Sources

AWS Glue Data Catalog
AWS Glue Data Catalog table as the data source.

Amazon S3
JSON, CSV, or Parquet files stored in S3.

Amazon Kinesis
Read from an Amazon Kinesis Data Stream.

Apache Kafka
Read from an Apache Kafka or Amazon MSK topic.

Relational DB
AWS Glue Data Catalog table with a relational database as the data source.

Amazon Redshift
Read your data from Amazon Redshift.

MySQL
AWS Glue Data Catalog table with MySQL as the data source.

PostgreSQL
AWS Glue Data Catalog table with PostgreSQL as the data source.

Oracle SQL
AWS Glue Data Catalog table with Oracle SQL as the data source.

Microsoft SQL Server
AWS Glue Data Catalog table with SQL Server as the data source.

Amazon DynamoDB
AWS Glue Data Catalog table with DynamoDB as the data source.

Snowflake
Read your data from Snowflake.

Popular transforms & data

At the top of the panel is a collection of **Popular transforms & data**. These nodes are commonly used in AWS Glue. Choose one to add it to the canvas. You can also hide the **Popular transforms & data** by clicking the triangle next to the **Popular transforms & data** heading.

Beneath the **Popular transforms & data** section, you can search for transforms and data source nodes. Results appear as you type. The more letters you add to your search query, the list of results will get smaller. Search results are populated from the node name and/or description. Choose the node to add it to your canvas.

Transforms and Data

There are two tabs that organize the nodes into **Transforms** and **Data**.

Transforms – When you choose the **Transforms** tab, all of the available transforms can be selected. Choose a transform to add it to the canvas. You can also choose **Add Transform** at the bottom of the Transforms list which will open a new page to the documentation for creating Custom visual transforms. Following the steps will allow you to create transforms of your own. Your transforms will then appear in the list of available transforms.

Data – The data tab contains all of the nodes for **Sources** and **Targets**. You can hide the Sources and Targets by clicking the triangle next to the Sources or Targets heading. You can unhide the Sources and Targets by clicking the triangle again. Choose a source or target node to add it to the canvas. You can also choose **Manage Connections** to add a new connection. This will open the Connectors page in the console.

AWS Glue for Spark and AWS Glue for Ray

In AWS Glue on Apache Spark (AWS Glue ETL), you can use PySpark to write Python code to handle data at scale. Spark is a familiar solution for this problem, but data engineers with Python-focused backgrounds can find the transition unintuitive. The Spark DataFrame model is not seamlessly "Pythonic", which reflects the Scala language and Java runtime it is built upon.

In AWS Glue, you can use Python shell jobs to run native Python data integrations. These jobs run on a single Amazon EC2 instance and are limited by the capacity of that instance. This restricts the throughput of the data you can process, and becomes expensive to maintain when dealing with big data.

AWS Glue for Ray allows you to scale up Python workloads without substantial investment into learning Spark. You can take advantage of certain scenarios where Ray performs better. By offering you a choice, you can use the strengths of both Spark and Ray.

AWS Glue ETL and AWS Glue for Ray are different underneath, so they support different features. Please check the documentation to determine supported features.

What is AWS Glue for Ray?

Ray is an open-source distributed computation framework that you can use to scale up workloads, with a focus on Python. For more information about Ray, see the Ray website. AWS Glue Ray jobs and interactive sessions allow you to use Ray within AWS Glue.

You can use AWS Glue for Ray to write Python scripts for computations that will run in parallel across multiple machines. In Ray jobs and interactive sessions, you can use familiar Python libraries, like pandas, to make your workflows easy to write and run. For more information about Ray datasets, see Ray Datasets in the Ray documentation. For more information about pandas, see the Pandas website.

When you use AWS Glue for Ray, you can run your pandas workflows against big data at enterprise scale—with only a few lines of code. You can create a Ray job from the AWS Glue console or the AWS SDK. You can also open an AWS Glue interactive session to run your code on a serverless Ray environment. Visual jobs in AWS Glue Studio are not yet supported.

AWS Glue for Ray jobs allow you to run a script on a schedule or in response to an event from Amazon EventBridge. Jobs store log information and monitoring statistics in CloudWatch that enable you to understand the health and reliability of your script. For more information about the AWS Glue job system, see the section called "Working with Ray jobs".

Ray automates the work of scaling Python code by distributing the processing across a cluster of machines that it reconfigures in real time, based on the load. This can lead to improved performance per dollar for certain workloads. With Ray jobs, we have built auto scaling natively into the AWS Glue job model, so you can fully take advantage of this feature. Ray jobs run on AWS Graviton, leading to higher overall price performance.

In addition to cost savings, you can use native auto scaling to run Ray workloads without investing time into cluster maintenance, tuning, and administration. You can use familiar open-source libraries out of the box, such as pandas, and the AWS SDK for Pandas. These improve iteration speed while you're developing on AWS Glue for Ray. When you use AWS Glue for Ray, you will be able to rapidly develop and run cost-effective data integration workloads.

Converting semi-structured schemas to relational schemas with AWS Glue

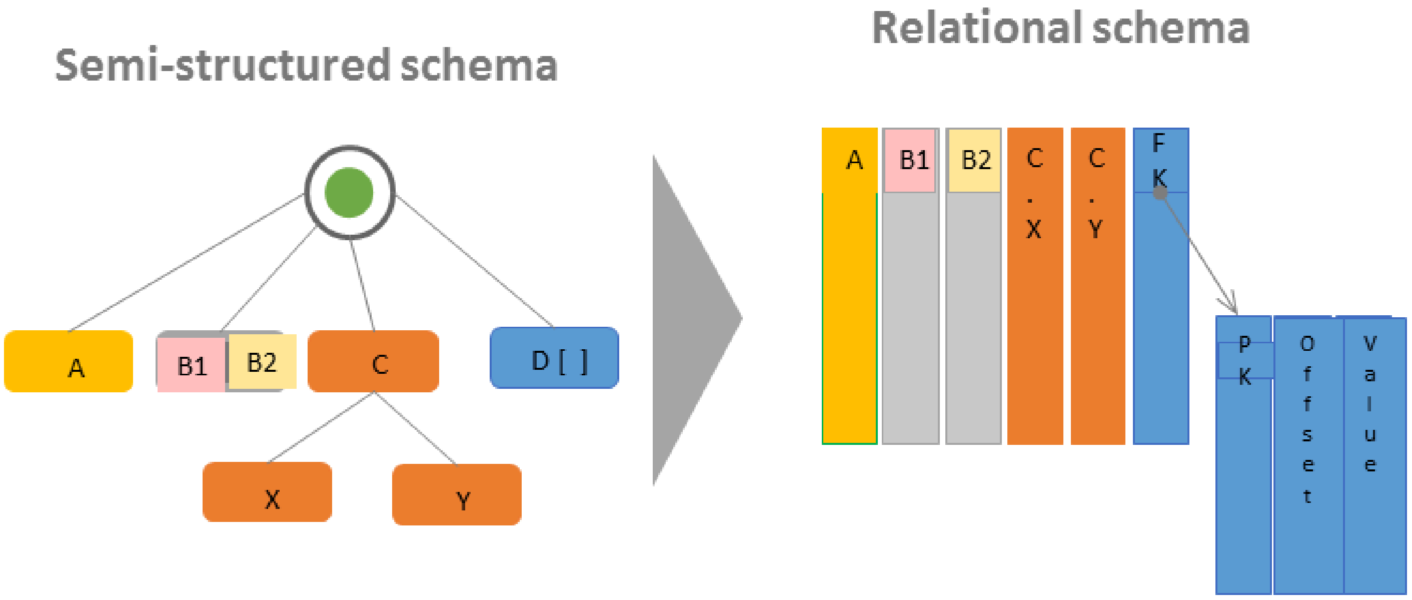
It's common to want to convert semi-structured data into relational tables. Conceptually, you are flattening a hierarchical schema to a relational schema. AWS Glue can perform this conversion for you on-the-fly.

Semi-structured data typically contains mark-up to identify entities within the data. It can have nested data structures with no fixed schema. For more information about semi-structured data, see Semi-structured data in Wikipedia.

Relational data is represented by tables that consist of rows and columns. Relationships between tables can be represented by a primary key (PK) to foreign key (FK) relationship. For more information, see Relational database in Wikipedia.

AWS Glue uses crawlers to infer schemas for semi-structured data. It then transforms the data to a relational schema using an ETL (extract, transform, and load) job. For example, you might want to parse JSON data from Amazon Simple Storage Service (Amazon S3) source files to Amazon Relational Database Service (Amazon RDS) tables. Understanding how AWS Glue handles the differences between schemas can help you understand the transformation process.

This diagram shows how AWS Glue transforms a semi-structured schema to a relational schema.


Semi-structured schema
Relational schema
A
B1
B2
C
D [ ]
X
Y
C.X
C.Y
FK
PK
Offset
Value

The diagram illustrates the following:

- Single value `A` converts directly to a relational column.
- The pair of values, `B1` and `B2`, convert to two relational columns.
- Structure `C`, with children `X` and `Y`, converts to two relational columns.
- Array `D[]` converts to a relational column with a foreign key (FK) that points to another relational table. Along with a primary key (PK), the second relational table has columns that contain the offset and value of the items in the array.

AWS Glue type systems

AWS Glue uses multiple type systems to provide a versatile interface over data systems that store data in very different ways. This document disambiguates AWS Glue type systems and data standards.

AWS Glue Data Catalog Types

The Data Catalog is a registry of tables and fields stored in various data systems, a metastore. When AWS Glue components, such as AWS Glue crawlers and AWS Glue with Spark jobs, write to the Data Catalog, they do so with an internal type system for tracking the types of fields. These values are shown in the **Data type** column of the table schema in the AWS Glue Console. This type system is based on Apache Hive's type system. For more information about the Apache Hive type system, see Types in the Apache Hive wiki. For more information about specific types and support, examples are provided in the AWS Glue Console, as part of the Schema Builder.

Validation, compatibility and other uses

The Data Catalog does not validate types written to type fields. When AWS Glue components read and write to the Data Catalog, they will be compatible with each other. AWS Glue components also aim to preserve a high degree of compatibility with the Hive types. However, AWS Glue components do not guarantee compatibility with all Hive types. This allows for interoperability with tools like Athena DDL when working with tables in the Data Catalog.

Since the Data Catalog does not validate types, other services may use the Data Catalog to track types using systems that strictly conform to the Hive type system, or any other system.

Types in AWS Glue with Spark scripts

When a AWS Glue with Spark script interprets or transforms a dataset, we provide `DynamicFrame`, an in-memory representation of your dataset as it is used in your script. The goal of a `DynamicFrame` is similar to that of the Spark `DataFrame`– it models your dataset so that Spark can schedule and execute transforms on your data. We guarantee that the type representation of `DynamicFrame` is intercompatible with `DataFrame` by providing the `toDF` and `fromDF` methods.

If type information can be inferred or provided to a `DataFrame`, it can be inferred or provided to a `DynamicFrame`, unless otherwise documented. When we provide optimized readers or writers for specific data formats, if Spark can read or write your data, our provided readers and writers will be able to, subject to documented limitations. For more information about readers and writers, see the section called “Data format options”.

The Choice Type

`DynamicFrames` provide a mechanism for modeling fields in a dataset whose value may have inconsistent types on disk across rows. For instance, a field may hold a number stored as a string in certain rows, and an integer in others. This mechanism is an in-memory type called `Choice`. We provide transforms such as the `ResolveChoice` method, to resolve Choice columns to a concrete type. AWS Glue ETL will not write the Choice type to the Data Catalog in the normal course of operation; Choice types only exist in the context of DynamicFrame memory models of datasets. For an example of Choice type usage, see the section called “Data preparation sample”.

AWS Glue Crawler Types

Crawlers aim to produce a consistent, usable schema for your dataset, then store it in Data Catalog for use in other AWS Glue components and Athena. Crawlers deal with types as described in the previous section on the Data Catalog, the section called “AWS Glue Data Catalog Types”. To produce a usable type in "Choice" type scenarios, where a column contains values of two or more types, Crawlers will create a `struct` type that models the potential types.

Getting started with AWS Glue

The following sections provide information on setting up AWS Glue. Not all of the setting up sections are required to start using AWS Glue. You can use the instructions as needed to set up IAM permissions, encryption, and DNS (if you're using a VPC environment to access data stores or if you're using interactive sessions).

Topics

- Overview of using AWS Glue
- Setting up IAM permissions for AWS Glue
- Setting up AWS Glue usage profiles
- Getting started with the AWS Glue Data Catalog
- Setting up network access to data stores
- Setting up encryption in AWS Glue
- Setting up networking for development for AWS Glue

Overview of using AWS Glue

With AWS Glue, you store metadata in the AWS Glue Data Catalog. You use this metadata to orchestrate ETL jobs that transform data sources and load your data warehouse or data lake. The following steps describe the general workflow and some of the choices that you make when working with AWS Glue.

> **Note**
>
> You can use the following steps, or you can create a workflow that automatically performs steps 1 through 3. For more information, see the section called “Performing complex ETL activities using blueprints and workflows”.

1. Populate the AWS Glue Data Catalog with table definitions.

   In the console, for persistent data stores, you can add a crawler to populate the AWS Glue Data Catalog. You can start the **Add crawler** wizard from the list of tables or the list of crawlers. You choose one or more data stores for your crawler to access. You can also create a schedule to determine the frequency of running your crawler. For data streams, you can manually create the table definition, and define stream properties.

Optionally, you can provide a custom classifier that infers the schema of your data. You can create custom classifiers using a grok pattern. However, AWS Glue provides built-in classifiers that are automatically used by crawlers if a custom classifier does not recognize your data. When you define a crawler, you don't have to select a classifier. For more information about classifiers in AWS Glue, see Defining and managing classifiers.

Crawling some types of data stores requires a connection that provides authentication and location information. If needed, you can create a connection that provides this required information in the AWS Glue console.

The crawler reads your data store and creates data definitions and named tables in the AWS Glue Data Catalog. These tables are organized into a database of your choosing. You can also populate the Data Catalog with manually created tables. With this method, you provide the schema and other metadata to create table definitions in the Data Catalog. Because this method can be a bit tedious and error prone, it's often better to have a crawler create the table definitions.

For more information about populating the AWS Glue Data Catalog with table definitions, see Creating tables.

2. Define a job that describes the transformation of data from source to target.

   Generally, to create a job, you have to make the following choices:

   - Choose a table from the AWS Glue Data Catalog to be the source of the job. Your job uses this table definition to access your data source and interpret the format of your data.
   - Choose a table or location from the AWS Glue Data Catalog to be the target of the job. Your job uses this information to access your data store.
   - Tell AWS Glue to generate a script to transform your source to target. AWS Glue generates the code to call built-in transforms to convert data from its source schema to target schema format. These transforms perform operations such as copy data, rename columns, and filter data to transform data as necessary. You can modify this script in the AWS Glue console.

   For more information about defining jobs in AWS Glue, see Building visual ETL jobs.

3. Run your job to transform your data.

   You can run your job on demand, or start it based on a one of these trigger types:

- A trigger that is based on a cron schedule.
- A trigger that is event-based; for example, the successful completion of another job can start an AWS Glue job.
- A trigger that starts a job on demand.

For more information about triggers in AWS Glue, see Starting jobs and crawlers using triggers.

4. Monitor your scheduled crawlers and triggered jobs.

   Use the AWS Glue console to view the following:

   - Job run details and errors.
   - Crawler run details and errors.
   - Any notifications about AWS Glue activities

   For more information about monitoring your crawlers and jobs in AWS Glue, see Monitoring AWS Glue.

Setting up IAM permissions for AWS Glue

The instructions in this topic help you quickly set up AWS Identity and Access Management (IAM) permissions for AWS Glue. You will complete the following tasks:

- Grant your IAM identities access to AWS Glue resources.
- Create a service role for running jobs, accessing data, and running AWS Glue Data Quality tasks.

For detailed instructions that you can use to customize IAM permissions for AWS Glue, see Configuring IAM permissions for AWS Glue.

To set up IAM permissions for AWS Glue in the AWS Management Console

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. Choose **Getting started**.
3. Under **Prepare your account for AWS Glue**, choose **Set up IAM permissions**.
4. Choose the IAM identities (roles or users) that you want to give AWS Glue permissions to. AWS Glue attaches the `AWSGlueConsoleFullAccess` managed policy to these identities. You

can skip this step if you want to set these permissions manually or only want to set a default service role.

5. Choose **Next**.
6. Choose the level of Amazon S3 access that your roles and users need. The options that you choose in this step are applied to all of the identities that you selected.

    a. Under **Choose S3 locations**, choose the Amazon S3 locations that you want to grant access to.

    b. Next, select whether your identities should have **Read only (recommended)** or **Read and write** access to the locations that you previously selected. AWS Glue adds permissions policies to your identities based on the combination of locations and read or write permissions you select.

    The following table displays the permissions that AWS Glue attaches for Amazon S3 access.

If you choose ...	AWS Glue attaches ...
No change	No permissions. AWS Glue won't make any changes to your identity's permissions.
Grant access to specific Amazon S3 locations (read only)	The customer-managed policy, `AWSGlueConsole-S3-read-only-policy` , grants access to specific Amazon S3 locations with read-only permissions. JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "s3:GetObject"
      ],
```

If you choose ...	AWS Glue attaches ...

```
      "Resource": [
        "arn:aws:s3:::amzn-s3-demo-bucket1/*",
        "arn:aws:s3:::amzn-s3-demo-bucket2/*",
        "arn:aws:s3:::amzn-s3-demo-bucket3",
        "arn:aws:s3:::amzn-s3-demo-bucket"
      ],
      "Condition": {
        "StringEquals": {
          "aws:ResourceAccount": "0000000000"
        }
      }
    }
  ]
}
```

If you choose ...	AWS Glue attaches ...
Grant access to specific Amazon S3 locations (read and write)	The `AWSGlueConsole-S3-read-and-write-policy` grants access to specific Amazon S3 locations with read and write permissions. JSON

```json
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "s3:GetObject",
        "s3:PutObject"
      ],
      "Resource": [
        "arn:aws:s3:::aes-siem-00000000000-log/*",
        "arn:aws:s3:::aes-siem-00000000000-snapshot/*",
        "arn:aws:s3:::aes-siem-00000000000-log",
        "arn:aws:s3:::aes-siem-00000000000-snapshot"
      ],
      "Condition": {
        "StringEquals": {
          "aws:ResourceAccount": "00000000000"
        }
      }
    }
  ]
}
```

7. Choose **Next**.

8. Choose a default AWS Glue service role for your account. A service role is an IAM role that AWS Glue uses to access resources in other AWS services on your behalf. For more information, see Service roles for AWS Glue.

    - When you choose the standard AWS Glue service role, AWS Glue creates a new IAM role in your AWS account named `AWSGlueServiceRole` with the following managed policies attached. If your account already has an IAM role named `AWSGlueServiceRole`, AWS Glue attaches these policies to the existing role.

        - AWSGlueServiceRole – This managed policy is required for AWS Glue to access and manage resources on your behalf. It allows AWS Glue to create, update, and delete various resources such as AWS Glue jobs, crawlers, and connections. This policy also grants permissions for AWS Glue to access Amazon CloudWatch logs for logging purposes. For the purposes of getting started, we recommend using this policy to learn how to use AWS Glue. As you get more comfortable with AWS Glue, you can create policies that allow you to fine-tune access to resources as needed.

        - AWSGlueConsoleFullAccess – This managed policy grants full access to the AWS Glue service through the AWS Management Console. This policy grants permissions to perform any operation within AWS Glue, enabling you to create, modify, and delete any AWS Glue resource as needed. However, it's important to note that this policy does not grant permissions to access the underlying data stores or other AWS services that may be involved in the ETL process. Due to the broad scope of permissions granted by the `AWSGlueConsoleFullAccess` policy, it should be assigned with caution and following the principle of least privilege. It is generally recommended to create and use more granular policies tailored to specific use cases and requirements whenever possible.

        - AWSGlueConsole-S3-read-only-policy – This policy allows AWS Glue to read data from specified Amazon S3 buckets, but it does not grant permissions to write or modify data in Amazon S3 or

          AWSGlueConsole-S3-read-and-write – This policy allows AWS Glue to read and write data to specified Amazon S3 buckets as part of the ETL process.

    - When you choose an existing IAM role, AWS Glue sets the role as the default, but doesn't add `AWSGlueServiceRole` permissions to it. Ensure that you've configured the role to use as a service role for AWS Glue. For more information, see Step 1: Create an IAM policy for the AWS Glue service and Step 2: Create an IAM role for AWS Glue.

9. Choose **Next**.

10. Finally, review the permissions you've selected and then choose **Apply changes**. When you apply the changes, AWS Glue adds IAM permissions to the identities that you selected. You can view or modify the new permissions in the IAM console at https://console.aws.amazon.com/iam/.

You've now completed the minimum IAM permissions setup for AWS Glue. In a production environment, we recommend that you familiarize yourself with Security in AWS Glue and Identity and access management for AWS Glue to help you secure AWS resources for your use case.

Next steps

Now that you have IAM permissions set up, you can explore the following topics to get started using AWS Glue:

- Getting Started with AWS Glue in AWS Skill Builder
- Getting started with the AWS Glue Data Catalog

Setting up for AWS Glue Studio

Complete the tasks in this section when you're using AWS Glue for the visual ETL for the first time:

Topics

- Review IAM permissions needed for the AWS Glue Studio user
- Review IAM permissions needed for ETL jobs
- Set up IAM permissions for AWS Glue Studio
- Configure a VPC for your ETL job

Review IAM permissions needed for the AWS Glue Studio user

To use AWS Glue Studio, the user must have access to various AWS resources. The user must be able to view and select Amazon S3 buckets, IAM policies and roles, and AWS Glue Data Catalog objects.

AWS Glue service permissions

AWS Glue Studio uses the actions and resources of the AWS Glue service. Your user needs permissions on these actions and resources to effectively use AWS Glue Studio. You can grant the

AWS Glue Studio user the `AWSGlueConsoleFullAccess` managed policy, or create a custom policy with a smaller set of permissions.

> ⚠ **Important**
>
> Per security best practices, it is recommended to restrict access by tightening policies to further restrict access to Amazon S3 bucket and Amazon CloudWatch log groups. For an example Amazon S3 policy, see Writing IAM Policies: How to Grant Access to an Amazon S3 Bucket.

Creating Custom IAM Policies for AWS Glue Studio

You can create a custom policy with a smaller set of permissions for AWS Glue Studio. The policy can grant permissions for a subset of objects or actions. Use the following information when creating a custom policy.

To use the AWS Glue Studio APIs, include `glue:UseGlueStudio` in the action policy in your IAM permissions. Using `glue:UseGlueStudio` will allow you to access all AWS Glue Studio actions even as more actions are added to the API over time.

For more information on actions defined by AWS Glue, see Actions defined by AWS Glue.

Data preparation authoring Actions

- SendRecipeAction
- GetRecipeAction

Directed acyclic graph (DAG) Actions

- CreateDag
- UpdateDag
- GetDag
- DeleteDag

Job Actions

- SaveJob

- GetJob
- CreateJob
- DeleteJob
- GetJobs
- UpdateJob

Job run Actions

- StartJobRun
- GetJobRuns
- BatchStopJobRun
- GetJobRun
- QueryJobRuns
- QueryJobs
- QueryJobRunsAggregated

Schema Actions

- GetSchema
- GetInferredSchema

Database Actions

- GetDatabases

Plan Actions

- GetPlan

Table Actions

- SearchTables
- GetTables
- GetTable

Connection Actions

- CreateConnection
- DeleteConnection
- UpdateConnection
- GetConnections
- GetConnection

Mapping Actions

- GetMapping

S3 Proxy Actions

- ListBuckets
- ListObjectsV2
- GetBucketLocation

Security Configuration Actions

- GetSecurityConfigurations

Script Actions

- CreateScript (different from API of same name in AWS Glue)

Accessing AWS Glue Studio APIs

To access AWS Glue Studio, add `glue:UseGlueStudio` in the actions policy list in the IAM permissions.

In the example below, `glue:UseGlueStudio` is included in the action policy, but the AWS Glue Studio APIs are not individually identified. That is because when you include `glue:UseGlueStudio`, you are automatically granted access to the internal APIs without having to specify the individual AWS Glue Studio APIs in the IAM permissions.

In the example, the additional listed action policies (for example, `glue:SearchTables`) are not AWS Glue Studio APIs, so they will need to be included in the IAM permissions as required. You may also want to include Amazon S3 Proxy actions to specify the level of Amazon S3 access to grant. The example policy below provides access to open AWS Glue Studio, create a visual job, and save/run it if the IAM role selected has sufficient access.

Notebook and data preview permissions

Data previews and notebooks allow you to see a sample of your data at any stage of your job (reading, transforming, writing), without having to run the job. You specify an AWS Identity and Access Management (IAM) role for AWS Glue Studio to use when accessing the data. IAM roles are intended to be assumable and do not have standard long-term credentials such as a password or access keys associated with it. Instead, when AWS Glue Studio assumes the role, IAM provides it with temporary security credentials.

To ensure data previews and notebook commands work correctly, use a role that has a name that starts with the string `AWSGlueServiceRole`. If you choose to use a different name for your role, then you must add the `iam:passrole` permission and configure a policy for the role in IAM. For more information, see Create an IAM policy for roles not named "AWSGlueServiceRole*".

> ⚠ **Warning**
>
> If a role grants the `iam:passrole` permission for a notebook, and you implement role chaining, a user could unintentionally gain access to the notebook. There is currently no auditing implemented which would allow you to monitor which users have been granted access to the notebook.

If you would like to deny an IAM identity the ability to create data preview sessions, consult the following example the section called “Deny an identity the ability to create data preview sessions”.

Amazon CloudWatch permissions

You can monitor your AWS Glue Studio jobs using Amazon CloudWatch, which collects and processes raw data from AWS Glue into readable, near-real-time metrics. By default, AWS Glue metrics data is sent to CloudWatch automatically. For more information, see What Is Amazon CloudWatch? in the *Amazon CloudWatch User Guide*, and AWS Glue Metrics in the *AWS Glue Developer Guide*.

To access CloudWatch dashboards, the user accessing AWS Glue Studio needs one of the following:

- The `AdministratorAccess` policy
- The `CloudWatchFullAccess` policy
- A custom policy that includes one or more of these specific permissions:
  - `cloudwatch:GetDashboard` and `cloudwatch:ListDashboards` to view dashboards
  - `cloudwatch:PutDashboard` to create or modify dashboards
  - `cloudwatch:DeleteDashboards` to delete dashboards

For more information for changing permissions for an IAM user using policies, see Changing Permissions for an IAM User in the *IAM User Guide*.

Review IAM permissions needed for ETL jobs

When you create a job using AWS Glue Studio, the job assumes the permissions of the IAM role that you specify when you create it. This IAM role must have permission to extract data from your data source, write data to your target, and access AWS Glue resources.

The name of the role that you create for the job must start with the string `AWSGlueServiceRole` for it to be used correctly by AWS Glue Studio. For example, you might name your role `AWSGlueServiceRole-FlightDataJob`.

Data source and data target permissions

An AWS Glue Studio job must have access to Amazon S3 for any sources, targets, scripts, and temporary directories that you use in your job. You can create a policy to provide fine-grained access to specific Amazon S3 resources.

- Data sources require `s3:ListBucket` and `s3:GetObject` permissions.
- Data targets require `s3:ListBucket`, `s3:PutObject`, and `s3:DeleteObject` permissions.

Note

Your IAM policy needs to allow `s3:GetObject` for the specific buckets used for hosting AWS Glue transforms.
The following buckets are owned by the AWS service account and is worldwide readable. These buckets serve as a repository for the source code pertinent to a subset of transformations accessible via the AWS Glue Studio visual editor. Permissions on the bucket are set up to deny any other API action on the bucket. Anybody can read those scripts we

provide for the transformations, but nobody outside our service team can "put" anything in them. When your AWS Glue job runs, that file is pulled in as a local import so the file is downloaded to the local container. After that, there is no further communication with that account.

Region: Bucket name

- af-south-1: aws-glue-studio-transforms-762339736633-prod-af-south-1
- ap-east-1: aws-glue-studio-transforms-125979764932-prod-ap-east-1
- ap-northeast-2: aws-glue-studio-transforms-673535381443-prod-ap-northeast-2
- ap-northeast-3: aws-glue-studio-transforms-149976050262-prod-ap-northeast-3
- ap-south-1: aws-glue-studio-transforms-584702181950-prod-ap-south-1
- ap-south-2: aws-glue-studio-transforms-380279651983-prod-ap-south-2
- ap-southeast-1: aws-glue-studio-transforms-737106620487-prod-ap-southeast-1
- ap-southeast-2: aws-glue-studio-transforms-234881715811-prod-ap-southeast-2
- ap-southeast-3: aws-glue-studio-transforms-151265630221-prod-ap-southeast-3
- ap-southeast-4: aws-glue-studio-transforms-052235663858-prod-ap-southeast-4
- ca-central-1: aws-glue-studio-transforms-622716468547-prod-ca-central-1
- ca-west-1: aws-glue-studio-transforms-915795495192-prod-ca-west-1
- eu-central-1: aws-glue-studio-transforms-560373232017-prod-eu-central-1
- eu-central-2: aws-glue-studio-transforms-907358657121-prod-eu-central-2
- eu-north-1: aws-glue-studio-transforms-312557305497-prod-eu-north-1
- eu-south-1: aws-glue-studio-transforms-939684186351-prod-eu-south-1
- eu-south-2: aws-glue-studio-transforms-239737454084-prod-eu-south-2
- eu-west-1: aws-glue-studio-transforms-244479516193-prod-eu-west-1
- eu-west-2: aws-glue-studio-transforms-804222392271-prod-eu-west-2
- eu-west-3: aws-glue-studio-transforms-371299348807-prod-eu-west-3
- il-central-1: aws-glue-studio-transforms-806964611811-prod-il-central-1
- me-central-1: aws-glue-studio-transforms-733304270342-prod-me-central-1
- me-south-1: aws-glue-studio-transforms-112120182341-prod-me-south-1

- sa-east-1: aws-glue-studio-transforms-881619130292-prod-sa-east-1
- us-east-1: aws-glue-studio-transforms-510798373988-prod-us-east-1
- us-east-2: aws-glue-studio-transforms-251189692203-prod-us-east-2
- us-west-1: aws-glue-studio-transforms-593230150239-prod-us-west-1
- us-west-2: aws-glue-studio-transforms-818035625594-prod-us-west-2
- ap-northeast-1: aws-glue-studio-transforms-200493242866-prod-ap-northeast-1
- cn-north-1: aws-glue-studio-transforms-071033555442-prod-cn-north-1
- cn-northwest-1: aws-glue-studio-transforms-070947029561-prod-cn-northwest-1
- us-gov-west-1: aws-glue-studio-transforms-227493901923-prod-us-gov-west-1-2604

If you choose Amazon Redshift as your data source, you can provide a role for cluster permissions. Jobs that run against a Amazon Redshift cluster issue commands that access Amazon S3 for temporary storage using temporary credentials. If your job runs for more than an hour, these credentials will expire causing the job to fail. To avoid this problem, you can assign a role to the Amazon Redshift cluster itself that grants the necessary permissions to jobs using temporary credentials. For more information, see Moving Data to and from Amazon Redshift in the *AWS Glue Developer Guide*.

If the job uses data sources or targets other than Amazon S3, then you must attach the necessary permissions to the IAM role used by the job to access these data sources and targets. For more information, see Setting Up Your Environment to Access Data Stores in the *AWS Glue Developer Guide*.

If you're using connectors and connections for your data store, you need additional permissions, as described in the section called “Permissions required for using connectors”.

Permissions required for deleting jobs

In AWS Glue Studio you can select multiple jobs in the console to delete. To perform this action, you must have the `glue:BatchDeleteJob` permission. This is different from the AWS Glue console, which requires the `glue:DeleteJob` permission for deleting jobs.

AWS Key Management Service permissions

If you plan to access Amazon S3 sources and targets that use server-side encryption with AWS Key Management Service (AWS KMS), then attach a policy to the AWS Glue Studio role used by the job that enables the job to decrypt the data. The job role needs the `kms:ReEncrypt`, `kms:GenerateDataKey`, and `kms:DescribeKey` permissions. Additionally, the job role needs the `kms:Decrypt` permission to upload or download an Amazon S3 object that is encrypted with an AWS KMS customer master key (CMK).

There are additional charges for using AWS KMS CMKs. For more information, see AWS Key Management Service Concepts - Customer Master Keys (CMKs) and AWS Key Management Service Pricing in the *AWS Key Management Service Developer Guide*.

Permissions required for using connectors

If you're using an AWS Glue Custom Connector and connection to access a data store, the role used to run the AWS Glue ETL job needs additional permissions attached:

- The AWS managed policy `AmazonEC2ContainerRegistryReadOnly` for accessing connectors purchased from AWS Marketplace.
- The `glue:GetJob` and `glue:GetJobs` permissions.
- AWS Secrets Manager permissions for accessing secrets that are used with connections. Refer to Example: Permission to retrieve secret values for example IAM policies.

If your AWS Glue ETL job runs within a VPC running Amazon VPC, then the VPC must be configured as described in the section called "Configure a VPC for your ETL job".

Set up IAM permissions for AWS Glue Studio

You can create the roles and assign policies to users and job roles by using the AWS administrator user.

You can use the **AWSGlueConsoleFullAccess** AWS managed policy to provide the necessary permissions for using the AWS Glue Studio console.

To create your own policy, follow the steps documented in Create an IAM Policy for the AWS Glue Service in the *AWS Glue Developer Guide*. Include the IAM permissions described previously in Review IAM permissions needed for the AWS Glue Studio user.

Topics

- Attach policies to the AWS Glue Studio user
- Create an IAM policy for roles not named "AWSGlueServiceRole*"

Attach policies to the AWS Glue Studio user

Any AWS user that signs in to the AWS Glue Studio console must have permissions to access specific resources. You provide those permissions by using assigning IAM policies to the user.

To attach the AWSGlueConsoleFullAccess managed policy to a user

1. Sign in to the AWS Management Console and open the IAM console at https://console.aws.amazon.com/iam/.
2. In the navigation pane, choose **Policies**.
3. In the list of policies, select the check box next to the **AWSGlueConsoleFullAccess**. You can use the **Filter** menu and the search box to filter the list of policies.
4. Choose **Policy actions**, and then choose **Attach**.
5. Choose the user to attach the policy to. You can use the **Filter** menu and the search box to filter the list of principal entities. After choosing the user to attach the policy to, choose **Attach policy**.
6. Repeat the previous steps to attach additional policies to the user, as needed.

Create an IAM policy for roles not named "AWSGlueServiceRole*"

To configure an IAM policy for roles used by AWS Glue Studio

1. Sign in to the AWS Management Console and open the IAM console at https://console.aws.amazon.com/iam/.
2. Add a new IAM policy. You can add to an existing policy or create a new IAM inline policy. To create an IAM policy:
   1. Choose **Policies**, and then choose **Create Policy**. If a **Get Started** button appears, choose it, and then choose **Create Policy**.
   2. Next to **Create Your Own Policy**, choose **Select**.
   3. For **Policy Name**, type any value that is easy for you to refer to later. Optionally, type descriptive text in **Description**.
   4. For **Policy Document**, type a policy statement with the following format, and then choose **Create Policy**:

3. Copy and paste the following blocks into the policy under the "Statement" array, replacing *my-interactive-session-role-prefix* with the prefix for all common roles to associate with permissions for AWS Glue.

   ```
   {
       "Action": [
           "iam:PassRole"
       ],
       "Effect": "Allow",
       "Resource": "arn:aws:iam::*:role/my-interactive-session-role-prefix*",
       "Condition": {
           "StringLike": {
               "iam:PassedToService": [
                   "glue.amazonaws.com "
               ]
           }
       }
   }
   ```

   Here is the full example with the Version and Statement arrays included in the policy

   JSON

   ```
   {
     "Version": "2012-10-17",
     "Statement": [
       {
         "Action": [
           "iam:PassRole"
         ],
         "Effect": "Allow",
         "Resource": "arn:aws:iam::*:role/my-interactive-session-role-prefix*",
         "Condition": {
           "StringLike": {
             "iam:PassedToService": [
               "glue.amazonaws.com "
             ]
           }
         }
       }
     ]
   ```

```
}
```

4. To enable the policy for a user, choose **Users**.
5. Choose the user to whom you want to attach the policy.

Configure a VPC for your ETL job

You can use Amazon Virtual Private Cloud (Amazon VPC) to define a virtual network in your own logically isolated area within the AWS Cloud, known as a *virtual private cloud (VPC)*. You can launch your AWS resources, such as instances, into your VPC. Your VPC closely resembles a traditional network that you might operate in your own data center, with the benefits of using the scalable infrastructure of AWS. You can configure your VPC; you can select its IP address range, create subnets, and configure route tables, network gateways, and security settings. You can connect instances in your VPC to the internet. You can connect your VPC to your own corporate data center, making the AWS Cloud an extension of your data center. To protect the resources in each subnet, you can use multiple layers of security, including security groups and network access control lists. For more information, see the Amazon VPC User Guide.

You can configure your AWS Glue ETL jobs to run within a VPC when using connectors. You must configure your VPC for the following, as needed:

- Public network access for data stores not in AWS. All data stores that are accessed by the job must be available from the VPC subnet.
- If your job needs to access both VPC resources and the public internet, the VPC needs to have a network address translation (NAT) gateway inside the VPC.

  For more information, see Setting Up Your Environment to Access Data Stores in the *AWS Glue Developer Guide*.

Getting started with notebooks in AWS Glue Studio

When you start a notebook through AWS Glue Studio, all the configuration steps are done for you so that you can explore your data and start developing your job script after only a few seconds.

The following sections describe how to create a role and grant the appropriate permissions to use notebooks in AWS Glue Studio for ETL jobs.

For more information on actions defined by AWS Glue, see Actions defined by AWS Glue .

Topics

- Granting permissions for the IAM role

Granting permissions for the IAM role

Setting up AWS Glue Studio is a pre-requisite to using notebooks.

To use notebooks in AWS Glue, your role requires the following:

- A trust relationship with AWS Glue for the `sts:AssumeRole` action and, if you want tagging then `sts:TagSession`.
- An IAM policy containing all the permissions for notebooks, AWS Glue, and interactive sessions.
- An IAM policy for a pass role since the role needs to be able to pass itself from the notebook to interactive sessions.

For example, when you create a new role, you can add a standard AWS managed policy like `AWSGlueConsoleFullAccessRole` to the role, and then add a new policy for the notebook operations and another for the IAM PassRole policy.

Actions needed for a trust relationship with AWS Glue

When starting a notebook session, you must add the `sts:AssumeRole` to the trust relationship of the role that is passed to the notebook. If your session includes tags, you must also pass the `sts:TagSession` action. Without these actions, the notebook session cannot start.

For example:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {
        "Service": "glue.amazonaws.com"
      },
      "Action": "sts:AssumeRole"
    }
  ]
}
```

Policies containing IAM permissions for notebooks

The following sample policy describes the required AWS IAM permissions for notebooks. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:StartNotebook",
        "glue:TerminateNotebook",
        "glue:GlueNotebookRefreshCredentials",
        "glue:DeregisterDataPreview",
        "glue:GetNotebookInstanceStatus",
        "glue:GlueNotebookAuthorize"
      ],
      "Resource": "*"
    }
  ]
}
```

You can use the following IAM policies to allow access to specific resources:

- *AwsGlueSessionUserRestrictedNotebookServiceRole*: Provides full access to all AWS Glue resources except for sessions. Allows users to create and use only the notebook sessions that are associated with the user. This policy also includes other permissions needed by AWS Glue to manage AWS Glue resources in other AWS services.
- *AwsGlueSessionUserRestrictedNotebookPolicy*: Provides permissions that allows users to create and use only the notebook sessions that are associated with the user. This policy also includes permissions to explicitly allow users to pass a restricted AWS Glue session role.

IAM policy to pass a role

When you create a notebook with a role, that role is then passed to interactive sessions so that the same role can be used in both places. As such, the `iam:PassRole` permission needs to be part of the role's policy.

Create a new policy for your role using the following example. Replace the account number with your own and the role name.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": "iam:PassRole",
      "Resource": "arn:aws:iam::111122223333:role/<role_name>"
    }
  ]
}
```

Setting up AWS Glue usage profiles

One of the main advantages of using a cloud platform is its flexibility. However, with this ease of creating compute resources comes a risk of spiraling cloud costs when left unmanaged and without guardrails. As a result, admins need to balance avoiding high infrastructure costs while at the same time allowing users to work without unnecessary friction.

With AWS Glue usage profiles, admins can create different profiles for various classes of users within the account, such as developers, testers, and product teams. Each profile is a unique set of parameters that can be assigned to different types of users. For example, developers may need more workers and can have a higher number of maximum workers while product teams may need fewer workers and a lower timeout or idle timeout value.

Example of jobs and job runs behavior

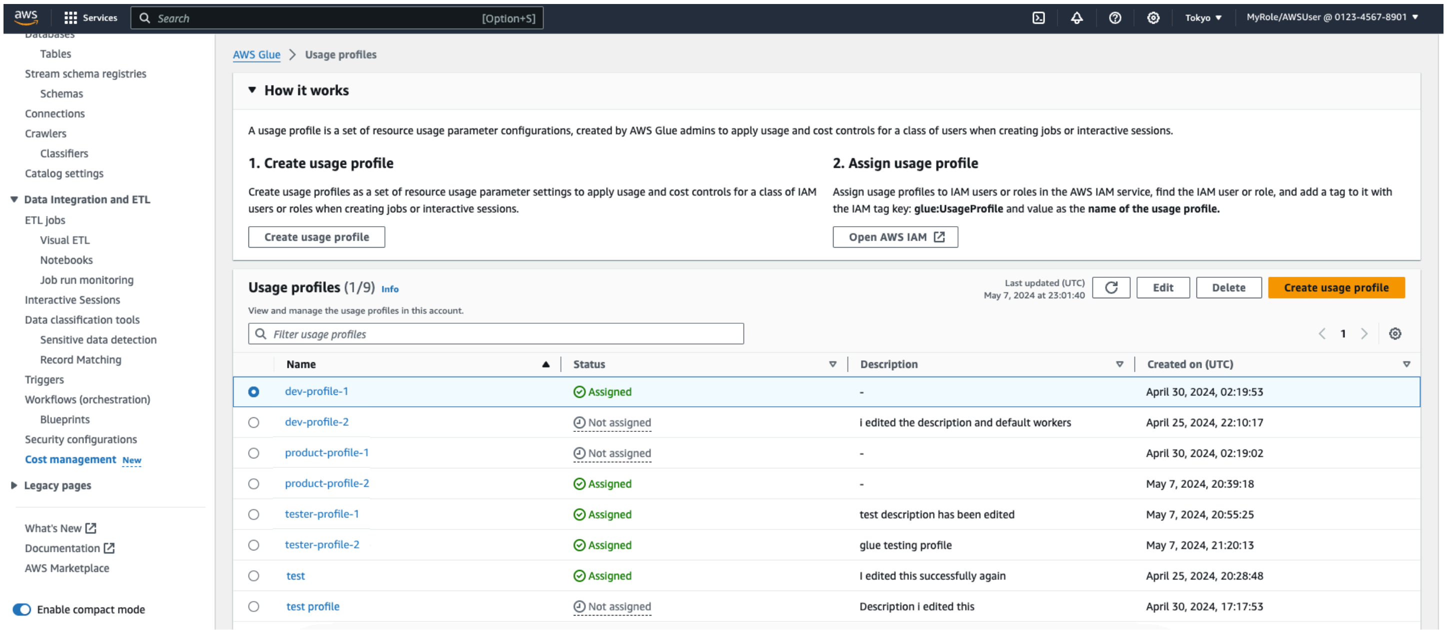
Suppose that a job is created by user A with profile A. The job is saved with certain parameter values. User B with profile B will try to run the job.

When user A authored the job, if he didn’t set a specific number of workers, the default set in user A's profile was applied and was saved with the job's definitions.

When user B runs the job, it run with whatever values were saved for it. If user B's own profile is more restrictive and not allowed to run with that many workers, the job run will fail.

Usage profile as a resource

An AWS Glue usage profile is a resource identified by an Amazon Resource Name (ARN). All the default IAM (Identity and Access Management) controls apply, including action-based and resource-based authorization. Admins should update the IAM policy of users who create AWS Glue resources, granting them access to use the profiles.


aws
Services
Search
[Option+S]
Tokyo ▾
MyRole/AWSUser @ 0123-4567-8901 ▾
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
▾ Data Integration and ETL
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Interactive Sessions
Data classification tools
Sensitive data detection
Record Matching
Triggers
Workflows (orchestration)
Blueprints
Security configurations
Cost management New
▸ Legacy pages
What's New
Documentation
AWS Marketplace
Enable compact mode
AWS Glue > Usage profiles
▾ How it works

A usage profile is a set of resource usage parameter configurations, created by AWS Glue admins to apply usage and cost controls for a class of users when creating jobs or interactive sessions.

1. Create usage profile

Create usage profiles as a set of resource usage parameter settings to apply usage and cost controls for a class of IAM users or roles when creating jobs or interactive sessions.

Create usage profile

2. Assign usage profile

Assign usage profiles to IAM users or roles in the AWS IAM service, find the IAM user or role, and add a tag to it with the IAM tag key: **glue:UsageProfile** and value as the **name of the usage profile.**

Open AWS IAM

Usage profiles (1/9) Info

Last updated (UTC)
May 7, 2024 at 23:01:40

Edit
Delete
Create usage profile

View and manage the usage profiles in this account.

Filter usage profiles
1

	Name	Status	Description	Created on (UTC)
●	dev-profile-1	✓ Assigned	-	April 30, 2024, 02:19:53
○	dev-profile-2	Not assigned	i edited the description and default workers	April 25, 2024, 22:10:17
○	product-profile-1	Not assigned	-	April 30, 2024, 02:19:02
○	product-profile-2	✓ Assigned	-	May 7, 2024, 20:39:18
○	tester-profile-1	✓ Assigned	test description has been edited	May 7, 2024, 20:55:25
○	tester-profile-2	✓ Assigned	glue testing profile	May 7, 2024, 21:20:13
○	test	✓ Assigned	I edited this successfully again	April 25, 2024, 20:28:48
○	test profile	Not assigned	Description i edited this	April 30, 2024, 17:17:53

Topics

- Creating and managing usage profiles
- Usage profiles and jobs

Creating and managing usage profiles

Creating an AWS Glue usage profile

Admins should create usage profiles and then assign them to the various users. When creating a usage profile, you specify default values as well as a range of allowed values for various job and session parameters. You must configure at least one parameter for jobs or interactive sessions. You can customize the default value to be used when a parameter value is not provided for the job, and/or set up a range limit or a set of allowed values for validation if a user provides a parameter value when using this profile.

Defaults are a best practice set by the admin to assist job authors. When a user creates a new job and doesn't set a timeout value, the usage profile's default timeout will apply. If the author doesn’t have a profile, then the AWS Glue service defaults would apply and be saved in the job's definition. At runtime, AWS Glue enforces the limits set in the profile (min, max, allowed workers).

Once a parameter is configured, all other parameters are optional. Parameters that can be customized for jobs or interactive sessions are:

- **Number of workers –** restrict the number of workers to avoid excessive use of compute resources. You can set a default, minimum, and maximum value. The minimum is 1.
- **Worker type –** restrict the relevant worker types for your workloads. You can set a default type and allow worker types for a user profile.
- **Timeout –** define the maximum time a job or interactive session can run and consume resources before it is terminated. Set up timeout values to avoid long-running jobs.

  You can set a default, minimum, and maximum value in minutes. The minimum is 1 (minute). While the AWS Glue default time out is 2880 minutes, you can set any default value in the usage profile.

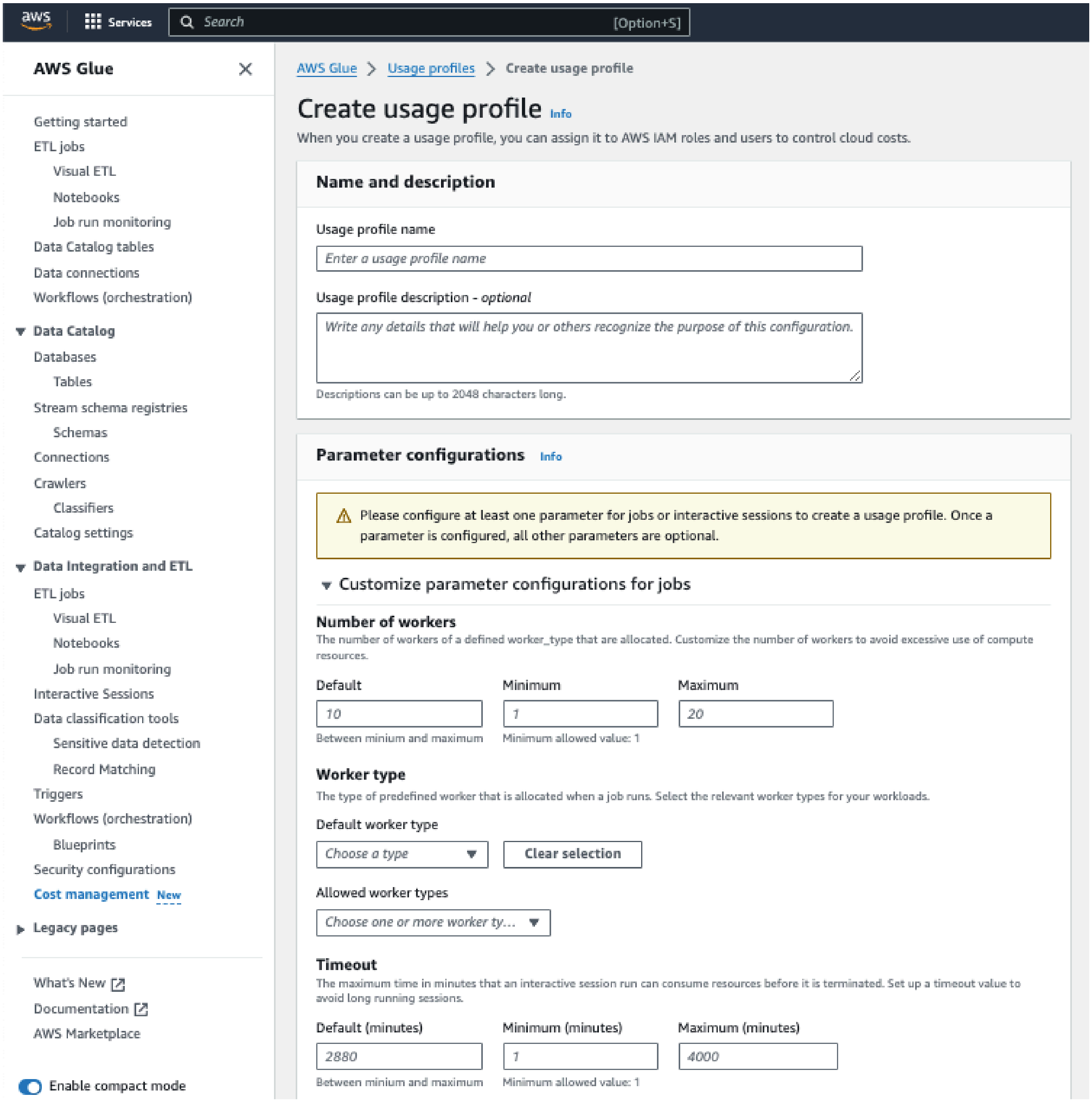
  It is a best practice to set a value for 'default'. This value will be used for the job or session creation if no value was set by the user.
- **Idle timeout –** define the number of minutes an interactive session is inactive before timing out after a cell has been run. Define idle timeout for interactive sessions to terminate after the work completed. Idle timeout range should be within the limit of timeout.

  You can set a default, minimum, and maximum value in minutes. The minimum is 1 (minute). While the AWS Glue default time out is 2880 minutes, you can set any default value in the usage profile.

  It is a best practice to set a value for 'default'. This value will be used for the session creation if no value was set by the user.

To create an AWS Glue usage profile as an admin (console)

1. In the left navigation menu, choose **Cost management**.
2. Choose **Create usage profile**.
3. Enter the **Usage profile name** for the usage profile.
4. Enter an optional description that will help others recognize the purpose of the usage profile.
5. Define at least one parameter in the profile. Any field in the form is a parameter. For example, the session idle timeout minimum.
6. Define any optional tags that apply to the usage profile.
7. Choose **Save**.


aws
Services
Search
[Option+S]
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Interactive Sessions
Data classification tools
Sensitive data detection
Record Matching
Triggers
Workflows (orchestration)
Blueprints
Security configurations
Cost management New
Legacy pages
What's New
Documentation
AWS Marketplace
Enable compact mode
AWS Glue > Usage profiles > Create usage profile
Info
Choose a type
Clear selection
Choose one or more worker ty...

Create usage profile

When you create a usage profile, you can assign it to AWS IAM roles and users to control cloud costs.

Name and description

Usage profile name: Enter a usage profile name

Usage profile description - *optional*: Write any details that will help you or others recognize the purpose of this configuration.

Descriptions can be up to 2048 characters long.

Parameter configurations

Please configure at least one parameter for jobs or interactive sessions to create a usage profile. Once a parameter is configured, all other parameters are optional.

Customize parameter configurations for jobs

Number of workers

The number of workers of a defined worker_type that are allocated. Customize the number of workers to avoid excessive use of compute resources.

Default	Minimum	Maximum
10	1	20
Between minium and maximum	Minimum allowed value: 1	

Worker type

The type of predefined worker that is allocated when a job runs. Select the relevant worker types for your workloads.

Default worker type

Allowed worker types

Timeout

The maximum time in minutes that an interactive session run can consume resources before it is terminated. Set up a timeout value to avoid long running sessions.

Default (minutes)	Minimum (minutes)	Maximum (minutes)
2880	1	4000
Between minium and maximum	Minimum allowed value: 1	

To create a usage profile (AWS CLI)

1. Enter the following command.

```
aws glue create-usage-profile --name profile-name --configuration file://
config.json --tags list-of-tags
```

where the config.json can define parameter values for interactive sessions (`SessionConfiguration`) and jobs (`JobConfiguration`):

```
//config.json (There is a separate blob for session/job configuration
{
    "SessionConfiguration": {
        "timeout": {
            "DefaultValue": "2880",
            "MinValue": "100",
            "MaxValue": "4000"
        },
        "idleTimeout": {
            "DefaultValue": "30",
            "MinValue": "10",
            "MaxValue": "4000"
        },
        "workerType": {
            "DefaultValue": "G.2X",
            "AllowedValues": [
                "G.1X",
                "G.2X",
                "G.4X",
                "G.8X",
                "G.12X",
                "G.16X",
                "R.1X",
                "R.2X",
                "R.4X",
                "R.8X"
            ]
        },
        "numberOfWorkers": {
            "DefaultValue": "10",
            "MinValue": "1",
            "MaxValue": "10"
        }
    },
    "JobConfiguration": {
        "timeout": {
            "DefaultValue": "2880",
            "MinValue": "100",
            "MaxValue": "4000"
        },
        "workerType": {
            "DefaultValue": "G.2X",
            "AllowedValues": [
                "G.1X",
                "G.2X",
                "G.4X",
                "G.8X",
                "G.12X",
                "G.16X",
                "R.1X",
                "R.2X",
                "R.4X",
                "R.8X"
            ]
        },
        "numberOfWorkers": {
            "DefaultValue": "10",
            "MinValue": "1",
            "MaxValue": "10"
        }
    }
}
```

2. Enter the following command to see the usage profile created:

```
aws glue get-usage-profile --name profile-name
```

The response:

```
{
    "ProfileName": "foo",
    "Configuration": {
        "SessionConfiguration": {
            "numberOfWorkers": {
                "DefaultValue": "10",
                "MinValue": "1",
                "MaxValue": "10"
            },
            "workerType": {
                "DefaultValue": "G.2X",
                "AllowedValues": [
                    "G.1X",
                    "G.2X",
                    "G.4X",
                    "G.8X",
                    "G.12X",
                    "G.16X",
                    "R.1X",
                    "R.2X",
                    "R.4X",
                    "R.8X"
                ]
            },
            "timeout": {
                "DefaultValue": "2880",
                "MinValue": "100",
                "MaxValue": "4000"
            },
            "idleTimeout": {
                "DefaultValue": "30",
                "MinValue": "10",
                "MaxValue": "4000"
            }
        },
        "JobConfiguration": {
            "numberOfWorkers": {
                "DefaultValue": "10",
                "MinValue": "1",
                "MaxValue": "10"
            },
            "workerType": {
                "DefaultValue": "G.2X",
                "AllowedValues": [
                    "G.1X",
                    "G.2X",
                    "G.4X",
                    "G.8X",
                    "G.12X",
                    "G.16X",
                    "R.1X",
                    "R.2X",
                    "R.4X",
                    "R.8X"
                ]
            },
            "timeout": {
                    "DefaultValue": "2880",
                    "MinValue": "100",
                    "MaxValue": "4000"
                }
            }
        },
        "CreatedOn": "2024-01-19T23:15:24.542000+00:00"
}
```

Additional CLI commands used to manage usage profiles:

- aws glue list-usage-profiles
- aws glue update-usage-profile --name *`profile-name`* --configuration *`file://config.json`*
- aws glue delete-usage-profile --name *`profile-name`*

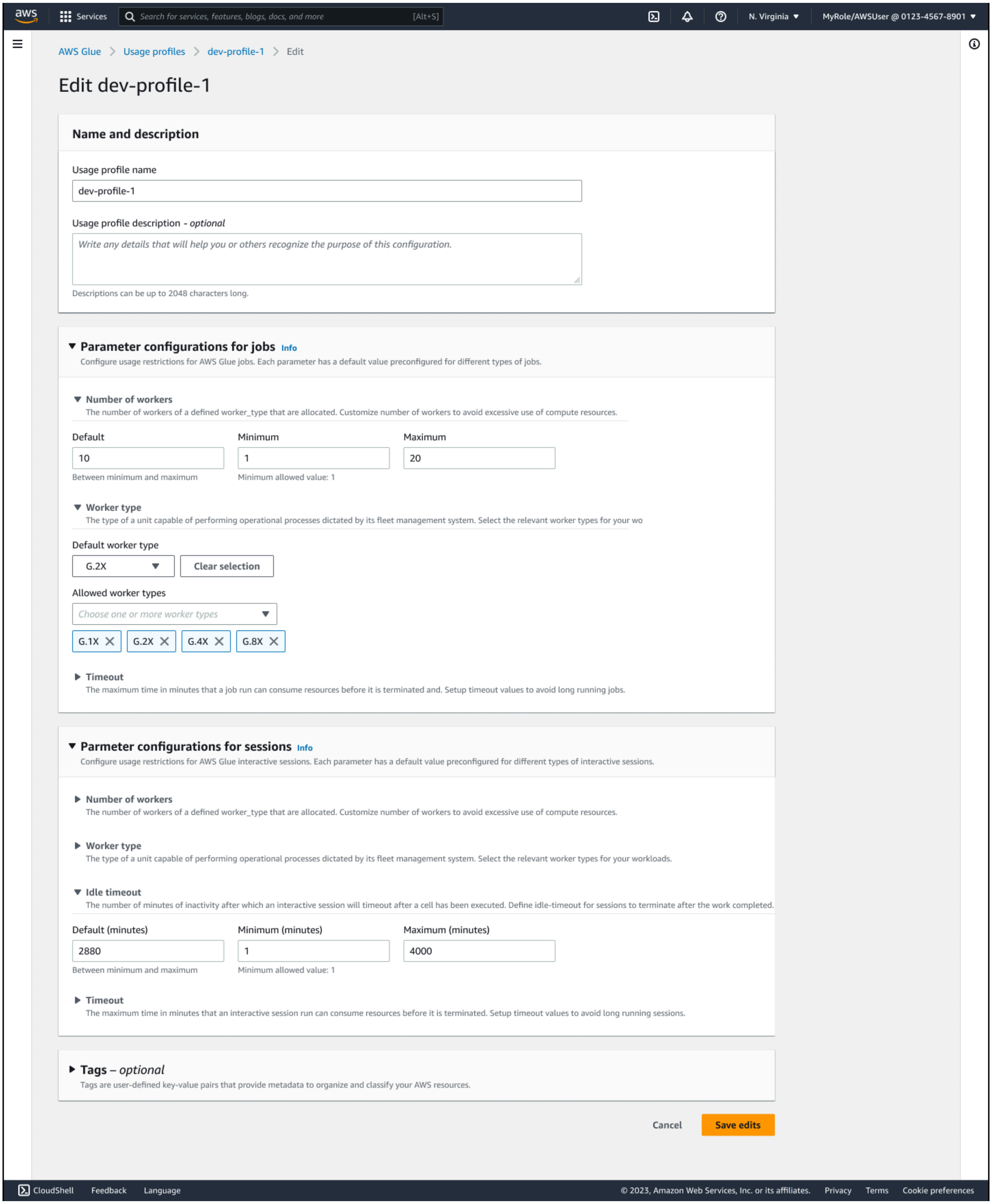
Editing a usage profile

Admins can edit usage profiles that they have created, to change the profile parameter values for jobs and interactive sessions.

To edit a usage profile:

To edit an AWS Glue usage profile as an admin (console)

1. In the left navigation menu, choose **Cost management**.
2. Choose a usage profile that you have permissions to edit and choose **Edit**.
3. Make changes as needed to the profile. By default, the parameters that already have values are expanded.
4. Choose **Save Edits**.


aws
Services
Search for services, features, blogs, docs, and more [Alt+S]
N. Virginia
MyRole/AWSUser @ 0123-4567-8901
AWS Glue > Usage profiles > dev-profile-1 > Edit

Edit dev-profile-1

Name and description

Usage profile name: dev-profile-1

Usage profile description - *optional*: *Write any details that will help you or others recognize the purpose of this configuration.*

Descriptions can be up to 2048 characters long.

Parameter configurations for jobs Info

Configure usage restrictions for AWS Glue jobs. Each parameter has a default value preconfigured for different types of jobs.

Number of workers

The number of workers of a defined worker_type that are allocated. Customize number of workers to avoid excessive use of compute resources.

Default	Minimum	Maximum
10	1	20
Between minimum and maximum	Minimum allowed value: 1	

Worker type

The type of a unit capable of performing operational processes dictated by its fleet management system. Select the relevant worker types for your wo

Default worker type: G.2X

Clear selection

Allowed worker types: *Choose one or more worker types*

G.1X ✕ G.2X ✕ G.4X ✕ G.8X ✕

Timeout

The maximum time in minutes that a job run can consume resources before it is terminated and. Setup timeout values to avoid long running jobs.

Parmeter configurations for sessions Info

Configure usage restrictions for AWS Glue interactive sessions. Each parameter has a default value preconfigured for different types of interactive sessions.

Number of workers

The number of workers of a defined worker_type that are allocated. Customize number of workers to avoid excessive use of compute resources.

Worker type

The type of a unit capable of performing operational processes dictated by its fleet management system. Select the relevant worker types for your workloads.

Idle timeout

The number of minutes of inactivity after which an interactive session will timeout after a cell has been executed. Define idle-timeout for sessions to terminate after the work completed.

Default (minutes)	Minimum (minutes)	Maximum (minutes)
2880	1	4000
Between minimum and maximum	Minimum allowed value: 1	

Timeout

The maximum time in minutes that an interactive session run can consume resources before it is terminated. Setup timeout values to avoid long running sessions.

Tags – *optional*

Tags are user-defined key-value pairs that provide metadata to organize and classify your AWS resources.

Cancel
Save edits
CloudShell
Feedback
Language
© 2023, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

To edit a usage profile (AWS CLI)

- Enter the following command. The same `--configuration` file syntax is used as shown above in the create command.

  ```
  aws glue update-usage-profile --name profile-name --configuration file://
  config.json
  ```

  where the config.json defines parameter values for interactive sessions (`SessionConfiguration`) and jobs (`JobConfiguration`):

Assigning a usage profile

The **Utilization status** column in the **Usage profiles** page shows whether a usage profile is assigned to users. Hovering over the status shows the assigned IAM entities.

The admin can assign an AWS Glue usage profile to users/roles who create AWS Glue resources. Assigning a profile is a combination of two actions:

- Updating the IAM user/role tag with the `glue:UsageProfile` key, then
- Updating the IAM policy of the user/role.

For users who use AWS Glue Studio to create jobs/interactive sessions, the admin tags the following roles:

- For restrictions on jobs, the admin tags the logged in console role
- For restrictions on interactive sessions, the admin tags the role the user provides when they create the notebook

The following is example policy that admin needs to update on the IAM users/roles who create AWS Glue resources:

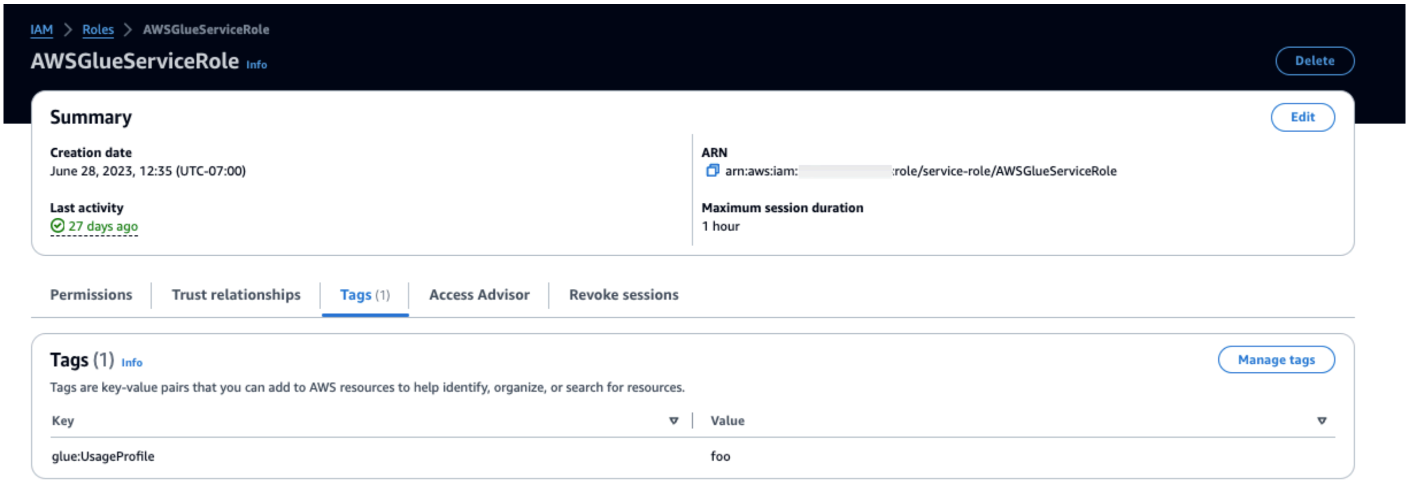
```
{
    "Effect": "Allow",
    "Action": [
        "glue:GetUsageProfile"
    ],
    "Resource": [
```

```
            "arn:aws:glue:us-east-1:123456789012:usageProfile/foo"
        ]
}
```

AWS Glue validates job, job run, and session requests based on the values specified in the AWS Glue usage profile and raises an exception if the request is disallowed. For synchronous APIs, an error will be thrown to the user. For asynchronous paths, a failed job run is created with the error message that the input parameter is outside of the allowed range for the assigned profile of the user/role.

To assign a usage profile to a user/role:

1. Open the (Identity and Access Management) IAM console.
2. In the left navigation, choose **Users** or **Roles**.
3. Choose a user or role.
4. Choose the **Tags** tab.
5. Choose **Add new tag**
6. Add a tag with the **Key** of `glue:UsageProfile` and the **Value** of the name of your usage profile.
7. Choose **Save changes**


IAM > Roles > AWSGlueServiceRole
AWSGlueServiceRole Info
Delete
Summary
Edit
Permissions
Trust relationships
Tags (1)
Access Advisor
Revoke sessions
Manage tags
Info

Creation date
June 28, 2023, 12:35 (UTC-07:00)

Last activity
27 days ago

ARN
arn:aws:iam:[redacted]:role/service-role/AWSGlueServiceRole

Maximum session duration
1 hour

Tags (1)

Tags are key-value pairs that you can add to AWS resources to help identify, organize, or search for resources.

Key	Value
glue:UsageProfile	foo

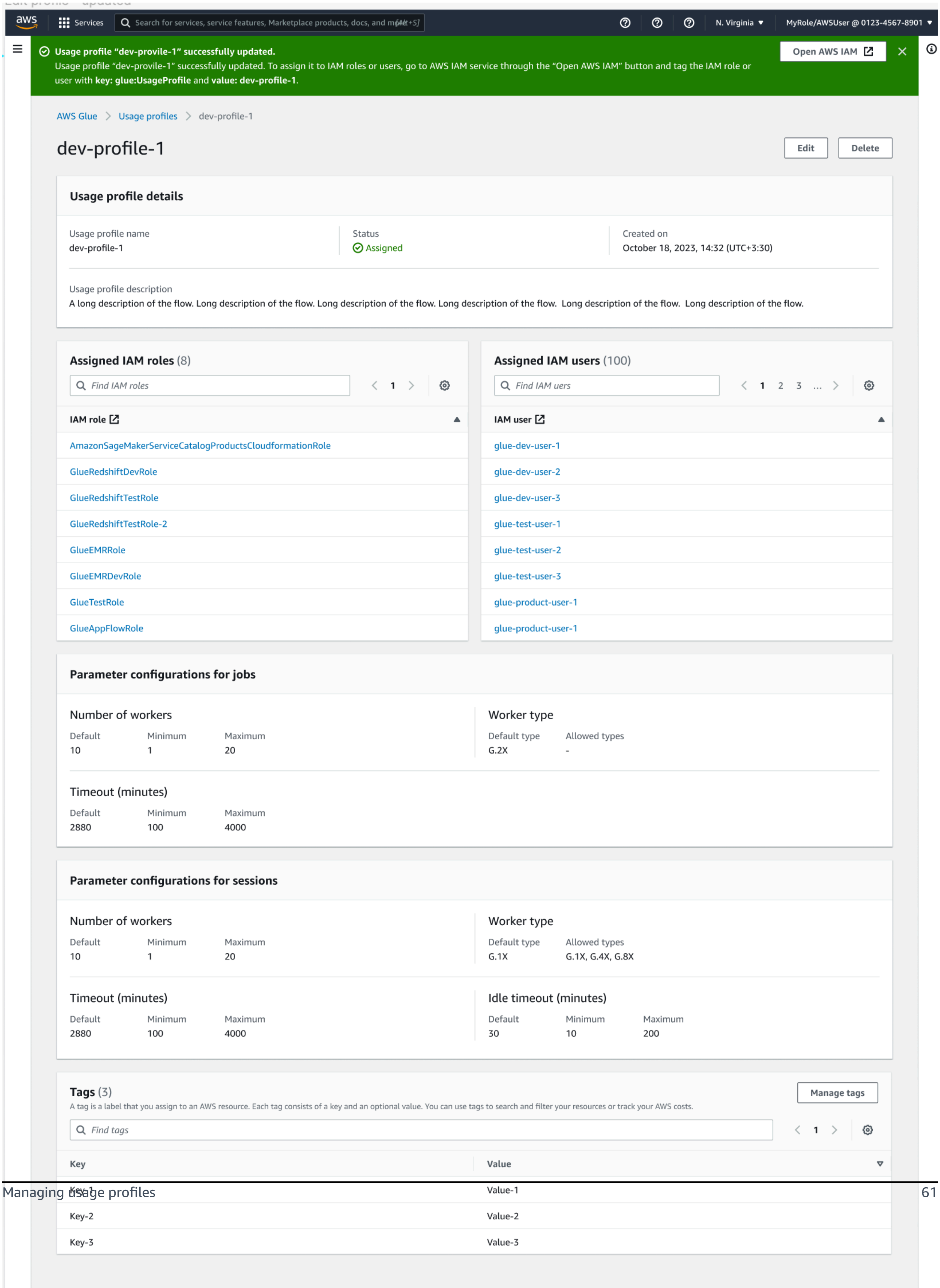
Viewing your assigned usage profile

Users can view their assigned usage profiles and use them when making API calls to create AWS Glue job and session resources, or starting a job.

Profile permissions are provided in IAM policies. As long as the caller policy has the `glue:UsageProfile` permission, a user can see the profile. Otherwise, you will get an access denied error.

To view an assigned usage profile:

1. In the left navigation menu, choose **Cost management**.
2. Choose a usage profile that you have permissions to view.


aws
Services
Search for services, features, Marketplace products, docs, and more [Alt+S]
N. Virginia
MyRole/AWSUser @ 0123-4567-8901
Open AWS IAM
Edit
Delete
Manage tags

Usage profile "dev-provile-1" successfully updated.
Usage profile "dev-provile-1" successfully updated. To assign it to IAM roles or users, go to AWS IAM service through the "Open AWS IAM" button and tag the IAM role or user with **key: glue:UsageProfile** and **value: dev-profile-1**.

AWS Glue > Usage profiles > dev-profile-1

dev-profile-1

Usage profile details

Usage profile name	Status	Created on
dev-profile-1	✓ Assigned	October 18, 2023, 14:32 (UTC+3:30)

Usage profile description
A long description of the flow. Long description of the flow. Long description of the flow. Long description of the flow. Long description of the flow. Long description of the flow.

Assigned IAM roles (8)

Find IAM roles

IAM role
AmazonSageMakerServiceCatalogProductsCloudformationRole
GlueRedshiftDevRole
GlueRedshiftTestRole
GlueRedshiftTestRole-2
GlueEMRRole
GlueEMRDevRole
GlueTestRole
GlueAppFlowRole

Assigned IAM users (100)

Find IAM uers

IAM user
glue-dev-user-1
glue-dev-user-2
glue-dev-user-3
glue-test-user-1
glue-test-user-2
glue-test-user-3
glue-product-user-1
glue-product-user-1

Parameter configurations for jobs

Number of workers

Default	Minimum	Maximum
10	1	20

Worker type

Default type	Allowed types
G.2X	-

Timeout (minutes)

Default	Minimum	Maximum
2880	100	4000

Parameter configurations for sessions

Number of workers

Default	Minimum	Maximum
10	1	20

Worker type

Default type	Allowed types
G.1X	G.1X, G.4X, G.8X

Timeout (minutes)

Default	Minimum	Maximum
2880	100	4000

Idle timeout (minutes)

Default	Minimum	Maximum
30	10	200

Tags (3)

A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs.

Find tags

Key	Value
Key-1	Value-1
Key-2	Value-2
Key-3	Value-3

Usage profiles and jobs

Authoring jobs with usage profiles

While authoring jobs, the limits and defaults set in your usage profile will apply. Your profile will be assigned to the job upon save.

Running jobs with usage profiles

When you start a job run, AWS Glue enforces the limits set in your caller's profile. If there is no direct caller, Glue will then apply the limits from the profile assigned to the job by its author.

> **Note**
>
> When a job is ran on a schedule (by AWS Glue workflows or AWS Glue triggers), the profile assigned to the job the author will apply.
> When a job is ran by an external service (Step Functions, MWAA) or a `StartJobRun` API, the caller's profile limit will be enforced.

For AWS Glue workflows or AWS Glue triggers: pre-existing jobs need to be updated to save the new profile name so that profile's limits (min, max, and allowed workers) will be enforced at runtime for scheduled runs.

Viewing a usage profile assigned for jobs

To view the profile assigned to your jobs (that will be used at runtime with scheduled AWS Glue workflows or AWS Glue triggers), you may look at the job **Details** tab. You may also look at the profile used in past runs in the job runs details tab.

Updating or deleting a usage profile attached to a job

The profile assigned to a job is changed upon update. If the author isn't assigned a usage profile, any profile previously attached to the job will be removed from it.

Getting started with the AWS Glue Data Catalog

The AWS Glue Data Catalog is your persistent technical metadata store. It is a managed service that you can use to store, annotate, and share metadata in the AWS Cloud. For more information, see AWS Glue Data Catalog.

The AWS Glue console and some user interfaces were recently updated.

Overview

You can use this tutorial to create your first AWS Glue Data Catalog, which uses an Amazon S3 bucket as your data source.

In this tutorial, you'll do the following using the AWS Glue console:

1. Create a database
2. Create a table
3. Use an Amazon S3 bucket as a data source

After completing these steps, you will have successfully used an Amazon S3 bucket as the data source to populate the AWS Glue Data Catalog.

Step 1: Create a database

To get started, sign in to the AWS Management Console and open the AWS Glue console.

To create a database using the AWS Glue console:

1. In the AWS Glue console, choose **Databases** under **Data catalog** from the left-hand menu.
2. Choose **Add database**.
3. In the Create a database page, enter a name for the database. In the **Location - *optional*** section, set the URI location for use by clients of the Data Catalog. If you don't know this, you can continue with creating the database.
4. (Optional). Enter a description for the database.
5. Choose **Create database**.

Congratulations, you've just set up your first database using the AWS Glue console. Your new database will appear in the list of available databases. You can edit the database by choosing the database's name from the **Databases** dashboard.

Next steps

Other ways to create a database:

You just created a database using the AWS Glue console, but there are other ways to create a database:

- You can use crawlers to create a database and tables for you automatically. To set up a database using crawlers, see Working with Crawlers in the AWS Glue Console.
- You can use AWS CloudFormation templates. See Creating AWS Glue Resources Using AWS Glue Data Catalog Templates.
- You can also create a database using the AWS Glue Database API operations.

  To create a database using the `create` operation, structure the request by including the `DatabaseInput` (required) parameters.

  For example:

  The following are examples of how you can use the CLI, Boto3, or DDL to define a table based on the same flights_data.csv file from the S3 bucket that you used in the tutorial.

  CLI

  ```
  aws glue create-database --database-input "{\"Name\":\"clidb\"}"
  ```

  Boto3

  ```
  glueClient = boto3.client('glue')

  response = glueClient.create_database(
      DatabaseInput={
          'Name': 'boto3db'
      }
  )
  ```

For more information about the Database API data types, structure, and operations, see Database API.

Next Steps

In the next section, you'll create a table and add that table to your database.

You can also explore the settings and permissions for your Data Catalog. See Working with Data Catalog Settings in the AWS Glue Console.

Step 2. Create a table

In this step, you create a table using the AWS Glue console.

1. In the AWS Glue console, choose **Tables** in the left-hand menu.
2. Choose **Add table**.
3. Set your table's properties by entering a name for your table in **Table details**.
4. In the **Databases** section, choose the database that you created in Step 1 from the drop-down menu.
5. In **Add a data store** section, **S3** will be selected by default as the type of source.
6. For **Data is located in** , choose **Specified path in another account**.
7. Copy and paste the path for the **Include path** input field:

   ```
   s3://crawler-public-us-west-2/flight/2016/csv/
   ```
8. In the section **Data format**, for **Classification**, choose **CSV**. and for **Delimiter**, choose **comma (,)**. Choose **Next**.
9. You are asked to define a schema. A schema defines the structure and format of a data record. Choose **Add column**. (For more information, see See Schema registries).
10. Specify the column properties:
    a. Enter a column name.
    b. For **Column type**, 'string' is already selected by default.
    c. For **Column number**, '1' is already selected by default.
    d. Choose **Add**.
11. You are asked to add partition indexes. This is optional. To skip this step, choose **Next**.
12. A summary of the table properties is displayed. If everything looks as expected, choose **Create**. Otherwise, choose **Back** and make edits as needed.

Congratulations, you've successfully created a table manually and associated it to a database. Your newly created table will appear in the Tables dashboard. From the dashboard, you can modify and manage all your tables.

For more information, see Working with Tables in the AWS Glue Console.

Next steps

Next steps

Now that the Data Catalog is populated, you can begin authoring jobs in AWS Glue. See Building visual ETL jobs with AWS Glue Studio.

In addition to using the console, there are other ways to define tables in the Data Catalog including:

- Creating and running a crawler
- Adding classifiers to a crawler in AWS Glue
- Using the AWS Glue Table API
- Using the AWS Glue Data Catalog template
- Migrating an Apache Hive metastore
- Using the AWS CLI, Boto3, or data definition language (DDL)

  The following are examples of how you can use the CLI, Boto3, or DDL to define a table based on the same flights_data.csv file from the S3 bucket that you used in the tutorial.

  See the documentation on how to structure an AWS CLI command. The CLI example contains the JSON syntax for the 'aws glue create-table --table-input' value.

  CLI

```
{
        "Name": "flights_data_cli",
        "StorageDescriptor": {
            "Columns": [
                {
                    "Name": "year",
                    "Type": "bigint"
                },
                {
                    "Name": "quarter",
                    "Type": "bigint"
                }
            ],
            "Location": "s3://crawler-public-us-west-2/flight/2016/csv",
            "InputFormat": "org.apache.hadoop.mapred.TextInputFormat",
            "OutputFormat":
 "org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat",
            "Compressed": false,
            "NumberOfBuckets": -1,
            "SerdeInfo": {
                "SerializationLibrary":
 "org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe",
                "Parameters": {
                    "field.delim": ",",
                    "serialization.format": ","
                }
            }
        },
        "PartitionKeys": [
            {
                "Name": "mon",
                "Type": "string"
            }
        ],
        "TableType": "EXTERNAL_TABLE",
        "Parameters": {
            "EXTERNAL": "TRUE",
            "classification": "csv",
            "columnsOrdered": "true",
            "compressionType": "none",
            "delimiter": ",",
            "skip.header.line.count": "1",
            "typeOfData": "file"
        }
    }
```

Boto3

```
import boto3

glue_client = boto3.client("glue")

response = glue_client.create_table(
    DatabaseName='sampledb',
    TableInput={
        'Name': 'flights_data_manual',
    'StorageDescriptor': {
      'Columns': [{
        'Name': 'year',
        'Type': 'bigint'
      }, {
        'Name': 'quarter',
        'Type': 'bigint'
      }],
      'Location': 's3://crawler-public-us-west-2/flight/2016/csv',
      'InputFormat': 'org.apache.hadoop.mapred.TextInputFormat',
      'OutputFormat':
 'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat',
      'Compressed': False,
      'NumberOfBuckets': -1,
      'SerdeInfo': {
        'SerializationLibrary':
 'org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe',
        'Parameters': {
          'field.delim': ',',
          'serialization.format': ','
        }
      },
    },
    'PartitionKeys': [{
      'Name': 'mon',
      'Type': 'string'
    }],
    'TableType': 'EXTERNAL_TABLE',
    'Parameters': {
      'EXTERNAL': 'TRUE',
      'classification': 'csv',
      'columnsOrdered': 'true',
      'compressionType': 'none',
      'delimiter': ',',
      'skip.header.line.count': '1',
      'typeOfData': 'file'
    }
    }
)
```

DDL

```
CREATE EXTERNAL TABLE `sampledb`.`flights_data` (
  `year` bigint,
  `quarter` bigint)
PARTITIONED BY (
  `mon` string)
ROW FORMAT DELIMITED
  FIELDS TERMINATED BY ','
STORED AS INPUTFORMAT
  'org.apache.hadoop.mapred.TextInputFormat'
OUTPUTFORMAT
  'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'
LOCATION
  's3://crawler-public-us-west-2/flight/2016/csv/'
TBLPROPERTIES (
  'classification'='csv',
  'columnsOrdered'='true',
  'compressionType'='none',
  'delimiter'=',',
  'skip.header.line.count'='1',
  'typeOfData'='file')
```

Setting up network access to data stores

To run your extract, transform, and load (ETL) jobs, AWS Glue must be able to access your data stores. If a job doesn't need to run in your virtual private cloud (VPC) subnet—for example, transforming data from Amazon S3 to Amazon S3—no additional configuration is needed.

If a job needs to run in your VPC subnet—for example, transforming data from a JDBC data store in a private subnet—AWS Glue sets up elastic network interfaces that enable your jobs to connect securely to other resources within your VPC. Each elastic network interface is assigned a private IP address from the IP address range within the subnet you specified. No public IP addresses are assigned. Security groups specified in the AWS Glue connection are applied on each of the elastic network interfaces. For more information, see Setting up Amazon VPC for JDBC connections to Amazon RDS data stores from AWS Glue.

All JDBC data stores that are accessed by the job must be available from the VPC subnet. To access Amazon S3 from within your VPC, a VPC endpoint is required. If your job needs to access both VPC resources and the public internet, the VPC needs to have a Network Address Translation (NAT) gateway inside the VPC.

A job or development endpoint can only access one VPC (and subnet) at a time. If you need to access data stores in different VPCs, you have the following options:

- Use VPC peering to access the data stores. For more about VPC peering, see VPC Peering Basics
- Use an Amazon S3 bucket as an intermediary storage location. Split the work into two jobs, with the Amazon S3 output of job 1 as the input to job 2.

For details on how to connect to a Amazon Redshift data store using Amazon VPC, see the section called “Configure Redshift”.

For details on how to connnect to Amazon RDS data stores using Amazon VPC, see the section called “Setting up Amazon VPC to connect to Amazon RDS data stores”.

Once necessary rules are set in Amazon VPC, you create a connection in AWS Glue with the necessary properties to connect to your data stores. For more information about the connection, see Connecting to data.

ⓘ **Note**

Make sure you set up your DNS environment for AWS Glue. For more information, see Setting up DNS in your VPC.

Topics

- Setting up a VPC to connect to PyPI for AWS Glue
- Setting up DNS in your VPC

Setting up a VPC to connect to PyPI for AWS Glue

The Python Package Index (PyPI) is a repository of software for the Python programming language. This topic addresses the details needed to support the use of pip installed packages (as specified by the session creator using the `--additional-python-modules` flag).

Using AWS Glue interactive sessions with a connector results in the use of VPC network via the subnet specified for the connector. Consequently AWS services and other network destinations are not available unless you set up a special configuration.

The resolutions to this issue include:

- Use of an internet gateway which is reachable by your session.
- Set up and use of an S3 bucket with a PyPI/simple repo containing the transitive closure of a package set's dependencies.
- Use of a CodeArtifact repository which is mirroring PyPI and attached to your VPC.

Setting up an internet gateway

The technical aspects are detailed in NAT gateway use cases but note these requirements for using `--additional-python-modules`. Specifically, `--additional-python-modules` requires access to pypi.org which is determined by the configuration of your VPC. Note the following requirements:

1. The requirement of installing additional python modules via pip install for a user's session. If the session uses a connector, your configuration may be affected.
2. When a connector is being used with `--additional-python-modules`, when the session is started the subnet associated with the connector's `PhysicalConnectionRequirements` has to provide a network path for reaching pypi.org.
3. You must determine whether or not your configuration is correct.

Setting up an Amazon S3 bucket to host a targeted PyPI/simple repo

This example sets up a PyPI mirror in Amazon S3 for a set of packages and their dependencies.

To set up the PyPI mirror for a set of packages:

```
# pip download all the dependencies
pip download -d s3pypi --only-binary :all: plotly gglplot
pip download -d s3pypi --platform manylinux_2_17_x86_64 --only-binary :all: psycopg2-binary
# create and upload the pypi/simple index and wheel files to the s3 bucket
s3pypi -b test-domain-name --put-root-index -v s3pypi/*
```

If you already have an existing artifact repository, it will have an index URL for pip's use that you can provide in place of the example URL for the Amazon S3 bucket as above.

To use the custom index-url, with some example packages:

```
%%configure
{
    "--additional-python-modules": "psycopg2_binary==2.9.5",
    "python-modules-installer-option": "--no-cache-dir --verbose --index-url https://
test-domain-name.s3.amazonaws.com/ --trusted-host test-domain-name.s3.amazonaws.com"
}
```

Setting up a CodeArtifact mirror of pypi attached to your VPC

To set up a mirror:

1. Create a repository in the same region as the subnet used by the connector.

   Select `Public upstream repositories` and choose `pypi-store`.
2. Provide access to the repository from the VPC for the subnet.
3. Specify the correct `--index-url` using the `python-modules-installer-option`.

   ```
   %%configure
   {
       "--additional-python-modules": "psycopg2_binary==2.9.5",
       "python-modules-installer-option": "--no-cache-dir --verbose --index-url https://
   test-domain-name.s3.amazonaws.com/ --trusted-host test-domain-name.s3.amazonaws.com"
   }
   ```

For more information, see Use CodeArtifact from a VPC.

Setting up DNS in your VPC

Domain Name System (DNS) is a standard by which names used on the internet are resolved to their corresponding IP addresses. A DNS hostname uniquely names a computer and consists of a host name and a domain name. DNS servers resolve DNS hostnames to their corresponding IP addresses.

To set up DNS in your VPC, ensure that DNS hostnames and DNS resolution are both enabled in your VPC. The VPC network attributes `enableDnsHostnames` and `enableDnsSupport` must be set to `true`. To view and modify these attributes, go to the VPC console at https://console.aws.amazon.com/vpc/.

For more information, see Using DNS with your VPC. Also, you can use the AWS CLI and call the modify-vpc-attribute command to configure the VPC network attributes.

> **Note**
>
> If you are using Route 53, confirm that your configuration does not override DNS network attributes.

Setting up encryption in AWS Glue

The following example workflow highlights the options to configure when you use encryption with AWS Glue. The example demonstrates the use of specific AWS Key Management Service (AWS KMS) keys, but you might choose other settings based on your particular needs. This workflow highlights only the options that pertain to encryption when setting up AWS Glue.

1. If the user of the AWS Glue console doesn't use a permissions policy that allows all AWS Glue API operations (for example, `"glue:*"`), confirm that the following actions are allowed:
   - `"glue:GetDataCatalogEncryptionSettings"`
   - `"glue:PutDataCatalogEncryptionSettings"`
   - `"glue:CreateSecurityConfiguration"`
   - `"glue:GetSecurityConfiguration"`
   - `"glue:GetSecurityConfigurations"`
   - `"glue:DeleteSecurityConfiguration"`
2. Any client that accesses or writes to an encrypted catalog—that is, any console user, crawler, job, or development endpoint—needs the following permissions.
3. Any user or role that accesses an encrypted connection password needs the following permissions.
4. The role of any extract, transform, and load (ETL) job that writes encrypted data to Amazon S3 needs the following permissions.

   JSON

   ```
   {
   ```

```
  "Version": "2012-10-17",
  "Statement": {
    "Effect": "Allow",
    "Action": [
      "kms:Decrypt",
      "kms:Encrypt",
      "kms:GenerateDataKey"
    ],
    "Resource": "arn:aws:kms:us-east-1:111122223333:key/key-id"
  }
}
```

5. Any ETL job or crawler that writes encrypted Amazon CloudWatch Logs requires the following permissions in the key and IAM policies.

   In the key policy (not the IAM policy):

```
{
  "Effect": "Allow",
  "Principal": {
   "Service": "logs.region.amazonaws.com"
  },
  "Action": [
   "kms:Encrypt*",
   "kms:Decrypt*",
   "kms:ReEncrypt*",
   "kms:GenerateDataKey*",
   "kms:Describe*"
  ],
  "Resource": "<arn of key used for ETL/crawler cloudwatch encryption>"
 }
```

   For more information about key policies, see Using Key Policies in AWS KMS in the *AWS Key Management Service Developer Guide*.

   In the IAM policy attach the `logs:AssociateKmsKey` permission:

```
{
  "Effect": "Allow",
  "Principal": {
   "Service": "logs.region.amazonaws.com"
  },
  "Action": [
   "logs:AssociateKmsKey"
  ],
  "Resource": "<arn of key used for ETL/crawler cloudwatch encryption>"
 }
```

6. Any ETL job that uses an encrypted job bookmark needs the following permissions.

   JSON

```
{
  "Version": "2012-10-17",
  "Statement": {
    "Effect": "Allow",
    "Action": [
      "kms:Decrypt",
      "kms:Encrypt"
    ],
    "Resource": "arn:aws:kms:us-east-1:111122223333:key/*"
  }
}
```

7. On the AWS Glue console, choose **Settings** in the navigation pane.
   a. On the **Data catalog settings** page, encrypt your Data Catalog by selecting **Metadata encryption**. This option encrypts all the objects in the Data Catalog with the AWS KMS key that you choose.
   b. For **AWS KMS key**, choose **aws/glue**. You can also choose a AWS KMS key that you created.

> ⚠ **Important**
>
> AWS Glue supports only symmetric customer master keys (CMKs). The **AWS KMS key** list displays only symmetric keys. However, if you select **Choose a AWS KMS key ARN**, the console lets you enter an ARN for any key type. Ensure that you enter only ARNs for symmetric keys.

When encryption is enabled, the client that is accessing the Data Catalog must have AWS KMS permissions.

8. In the navigation pane, choose **Security configurations**. A security configuration is a set of security properties that can be used to configure AWS Glue processes. Then choose **Add security configuration**. In the configuration, choose any of the following options:
    a. Select **S3 encryption**. For **Encryption mode**, choose **SSE-KMS**. For the **AWS KMS key**, choose **aws/s3** (ensure that the user has permission to use this key). This enables data written by the job to Amazon S3 to use the AWS managed AWS Glue AWS KMS key.
    b. Select **CloudWatch logs encryption**, and choose a CMK. (Ensure that the user has permission to use this key). For more information, see Encrypt Log Data in CloudWatch Logs Using AWS KMS in the *AWS Key Management Service Developer Guide*.

    > ⚠ **Important**
    >
    > AWS Glue supports only symmetric customer master keys (CMKs). The **AWS KMS key** list displays only symmetric keys. However, if you select **Choose a AWS KMS key ARN**, the console lets you enter an ARN for any key type. Ensure that you enter only ARNs for symmetric keys.

    c. Choose **Advanced properties**, and select **Job bookmark encryption**. For the **AWS KMS key**, choose **aws/glue** (ensure that the user has permission to use this key). This enables encryption of job bookmarks written to Amazon S3 with the AWS Glue AWS KMS key.
9. In the navigation pane, choose **Connections**.
    a. Choose **Add connection** to create a connection to the Java Database Connectivity (JDBC) data store that is the target of your ETL job.
    b. To enforce that Secure Sockets Layer (SSL) encryption is used, select **Require SSL connection**, and test your connection.
10. In the navigation pane, choose **Jobs**.
    a. Choose **Add job** to create a job that transforms data.
    b. In the job definition, choose the security configuration that you created.
11. On the AWS Glue console, run your job on demand. Verify that any Amazon S3 data written by the job, the CloudWatch Logs written by the job, and the job bookmarks are all encrypted.

Setting up networking for development for AWS Glue

To run your extract, transform, and load (ETL) scripts with AWS Glue, you can develop and test your scripts using a *development endpoint*. Development endpoints are not supported for use with

AWS Glue version 2.0 jobs. For versions 2.0 and later, the preferred development method is using Jupyter Notebook with one of the AWS Glue kernels. For more information, see the section called “Getting started with AWS Glue interactive sessions”.

Setting up your network for a development endpoint

When you set up a development endpoint, you specify a virtual private cloud (VPC), subnet, and security groups.

Note

Make sure you set up your DNS environment for AWS Glue. For more information, see Setting up DNS in your VPC.

To enable AWS Glue to access required resources, add a row in your subnet route table to associate a prefix list for Amazon S3 to the VPC endpoint. A prefix list ID is required for creating an outbound security group rule that allows traffic from a VPC to access an AWS service through a VPC endpoint. To ease connecting to a notebook server that is associated with this development endpoint, from your local machine, add a row to the route table to add an internet gateway ID. For more information, see VPC Endpoints. Update the subnet routes table to be similar to the following table:

Destination	Target		
10.0.0.0/16	local		
pl-id for Amazon S3	vpce-id		
0.0.0.0/0	igw-xxxx		

To enable AWS Glue to communicate between its components, specify a security group with a self-referencing inbound rule for all TCP ports. By creating a self-referencing rule, you can restrict the source to the same security group in the VPC, and it's not open to all networks. The default security group for your VPC might already have a self-referencing inbound rule for ALL Traffic.

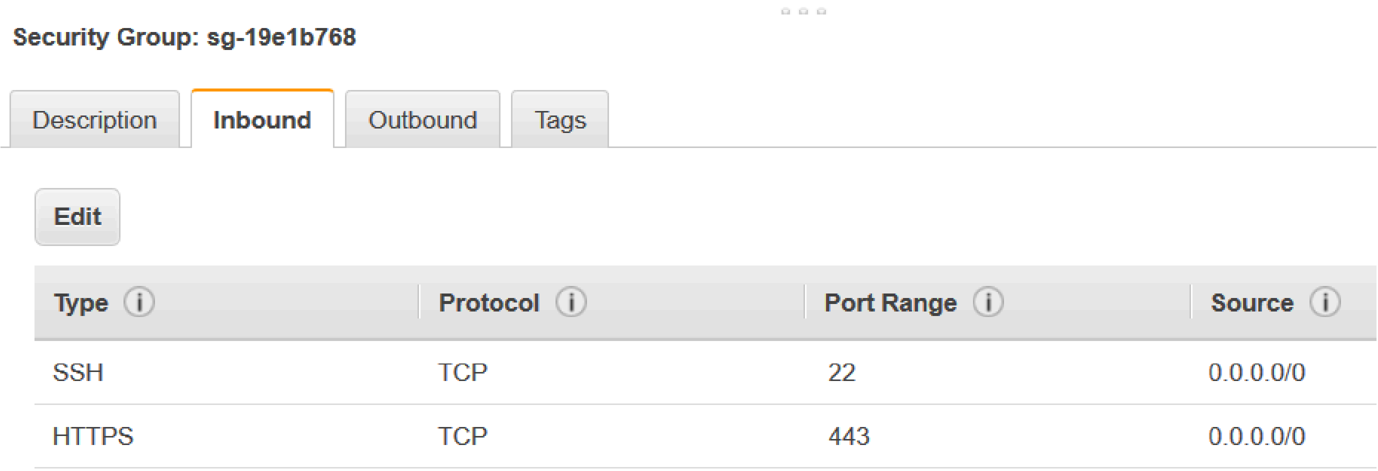
To set up a security group

1. Sign in to the AWS Management Console and open the Amazon EC2 console at https://console.aws.amazon.com/ec2/.
2. In the left navigation pane, choose **Security Groups**.
3. Either choose an existing security group from the list, or **Create Security Group** to use with the development endpoint.
4. In the security group pane, navigate to the **Inbound** tab.
5. Add a self-referencing rule to allow AWS Glue components to communicate. Specifically, add or confirm that there is a rule of **Type** `All TCP`, **Protocol** is `TCP`, **Port Range** includes all ports, and whose **Source** is the same security group name as the **Group ID**.

   The inbound rule looks similar to this:

Type	Protocol	Port range	Source
All TCP	TCP	0–65535	*`security-group`*

   The following shows an example of a self-referencing inbound rule:

   Security Group: sg-19e1b768

Type	Protocol	Port Range	Source
SSH	TCP	22	0.0.0.0/0
HTTPS	TCP	443	0.0.0.0/0

6. Add a rule to for outbound traffic also. Either open outbound traffic to all ports, or create a self-referencing rule of **Type** `All TCP`, **Protocol** is `TCP`, **Port Range** includes all ports, and whose **Source** is the same security group name as the **Group ID**.

   The outbound rule looks similar to one of these rules:

Type	Protocol	Port range	Destination
All TCP	TCP	0–65535	*security-group*
All Traffic	ALL	ALL	0.0.0.0/0

Setting up Amazon EC2 for a notebook server

With a development endpoint, you can create a notebook server to test your ETL scripts with Jupyter notebooks. To enable communication to your notebook, specify a security group with inbound rules for both HTTPS (port 443) and SSH (port 22). Ensure that the rule's source is either 0.0.0.0/0 or the IP address of the machine that is connecting to the notebook.

To set up a security group

1. Sign in to the AWS Management Console and open the Amazon EC2 console at https://console.aws.amazon.com/ec2/.
2. In the left navigation pane, choose **Security Groups**.
3. Either choose an existing security group from the list, or **Create Security Group** to use with your notebook server. The security group that is associated with your development endpoint is also used to create your notebook server.
4. In the security group pane, navigate to the **Inbound** tab.
5. Add inbound rules similar to this:

Type	Protocol	Port range	Source
SSH	TCP	22	0.0.0.0/0
HTTPS	TCP	443	0.0.0.0/0

The following shows an example of the inbound rules for the security group:

Data discovery and cataloging in AWS Glue

The AWS Glue Data Catalog is a centralized repository that stores metadata about your organization's data sets. It acts as an index to the location, schema, and runtime metrics of your data sources. The metadata is stored in metadata tables, where each table represents a single data store.

You can populate the Data Catalog using a crawler, which automatically scans your data sources and extracts metadata. A crawler can connect to data sources that are internal (AWS-based) and external to AWS.

For more information about the supported data sources, see Supported data sources for crawling

You can also create tables in the Data Catalog manually by defining the table structure, schema, and partitioning structure according to your specific requirements.

For more information about creating metadata tables manually, see Defining metadata manually.

You can use the information in the Data Catalog to create and monitor your ETL jobs. The Data Catalog integrates with other AWS analytics services, providing a unified view of data sources making it easier to manage and analyze data.

- Amazon Athena – Store and query table metadata in the Data Catalog for the Amazon S3 data using SQL.
- AWS Lake Formation – Centrally define and manage fine-grained data access policies and audit data access.
- Amazon EMR – Access data sources defined in the Data Catalog for big data processing.
- Amazon SageMaker AI – Quickly and confidently build, train, and deploy machine learning models.

Key features of the Data Catalog

The following are the key aspects of the Data Catalog.

Metadata repository

> The Data Catalog acts as a central metadata repository, storing information about the location, schema, and properties of your data sources. This metadata is organized into databases and tables, similar to a traditional relational database catalog.

Automatic data discoverability

AWS Glue crawlers can automatically discover and catalog new or updated data sources, reducing the overhead of manual metadata management and ensuring that your Data Catalog remains up-to-date. By cataloging your data sources, the Data Catalog makes it easier for users and applications to discover and understand the available data assets within your organization, promoting data reuse and collaboration.

The Data Catalog supports a wide range of data sources, including Amazon S3, Amazon RDS, Amazon Redshift, Apache Hive, and more. It can automatically infer and store metadata from these sources using AWS Glue crawlers.

For more information see, Using crawlers to populate the Data Catalog .

Schema management

The Data Catalog automatically captures and manages the schema of your data sources, including schema inference, evolution, and versioning. You can update your schema and partitions in the Data Catalog using AWS Glue ETL jobs.

Table optimization

For better read performance by AWS analytics services such as Amazon Athena and Amazon EMR, and AWS Glue ETL jobs, the Data Catalog provides managed compaction (a process that compacts small Amazon S3 objects into larger objects) for Iceberg tables in the Data Catalog. You can use AWS Glue console, AWS Lake Formation console, AWS CLI, or AWS API to enable or disable compaction for individual Iceberg tables that are in the Data Catalog.

For more information, see Optimizing Iceberg tables.

Column statistics

You can compute column-level statistics for Data Catalog tables in data formats such as Parquet, ORC, JSON, ION, CSV, and XML without setting up additional data pipelines. Column statistics help you to understand data profiles by getting insights about values within a column. The Data Catalog supports generating statistics for column values such as minimum value, maximum value, total null values, total distinct values, average length of values, and total occurrences of true values.

For more information, see Optimizing query performance using column statistics.

Data lineage

The Data Catalog maintains a record of the transformations and operations performed on your data, providing data lineage information. This lineage information is valuable for auditing, compliance, and understanding the data's provenance.

Integration with other AWS services

The Data Catalog seamlessly integrates with other AWS services, such as AWS Lake Formation, Amazon Athena, Amazon Redshift Spectrum, and Amazon EMR. This integration allows you to query and analyze data across various data stores using a single, consistent metadata layer.

Security and access control

AWS Glue integrates with AWS Lake Formation to support fine-grained access control for Data Catalog resources, allowing you to manage permissions and secure access to your data assets based on your organization's policies and requirements. AWS Glue integrates with AWS Key Management Service (AWS KMS) to encrypt metadata that's stored in the Data Catalog.

Topics

- Populating the AWS Glue Data Catalog
- Populating and managing transactional tables
- Managing the Data Catalog
- Accessing the Data Catalog
- AWS Glue Data Catalog best practices
- Monitoring Data Catalog usage metrics in Amazon CloudWatch
- AWS Glue Schema registry

Populating the AWS Glue Data Catalog

You can populate the AWS Glue Data Catalog using the following methods:

- AWS Glue crawler – An AWS Glue crawler can automatically discover and catalog data sources like databases, data lakes, and streaming data. The crawlers are the most common and recommended method to populate the Data Catalog as they can automatically discover and infer metadata for a wide variety of data sources.

- Manually adding metadata – You can manually define databases, tables, and connection details and add them to the Data Catalog using the AWS Glue console, Lake Formation console, AWS CLI, or AWS Glue APIs. Manual entry is useful when you want to catalog data sources that cannot be crawled.
- Integrating with other AWS services – You can populate the Data Catalog with metadata from services like AWS Lake Formation and Amazon Athena. These services can discover and register data sources in the Data Catalog.
- Populating from an existing metadata repository – If you have an existing metadata store like Apache Hive Metastore, you can use AWS Glue to import that metadata into the Data Catalog. For more information, see Migration between the Hive Metastore and the AWS Glue Data Catalog on GitHub.

Topics

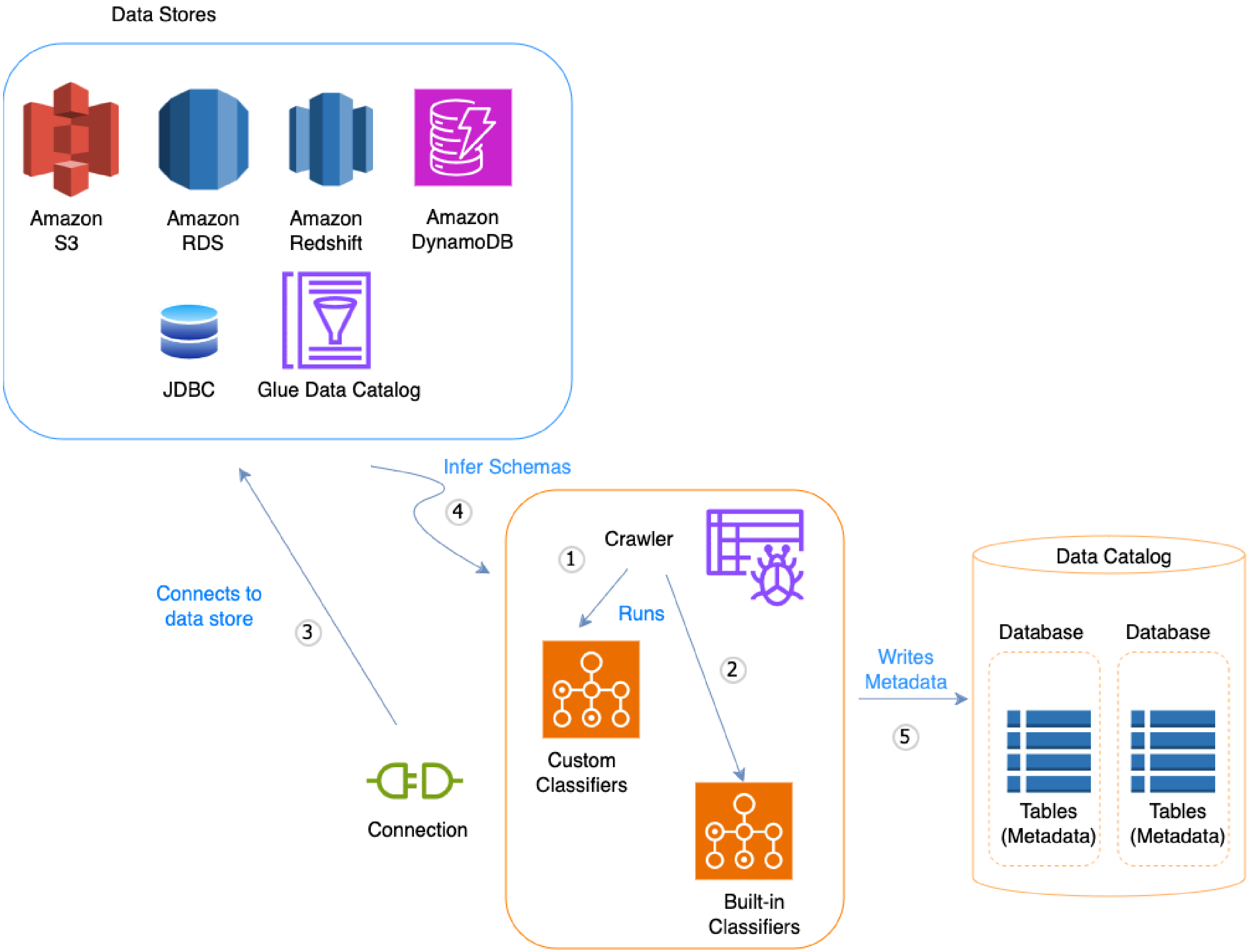
- Using crawlers to populate the Data Catalog
- Defining metadata manually
- Integrating with other AWS services
- Data Catalog settings

Using crawlers to populate the Data Catalog

You can use an AWS Glue crawler to populate the AWS Glue Data Catalog with databases and tables. This is the primary method used by most AWS Glue users. A crawler can crawl multiple data stores in a single run. Upon completion, the crawler creates or updates one or more tables in your Data Catalog. Extract, transform, and load (ETL) jobs that you define in AWS Glue use these Data Catalog tables as sources and targets. The ETL job reads from and writes to the data stores that are specified in the source and target Data Catalog tables.

Workflow

The following workflow diagram shows how AWS Glue crawlers interact with data stores and other elements to populate the Data Catalog.


Data Stores
Amazon S3
Amazon RDS
Amazon Redshift
Amazon DynamoDB
JDBC
Glue Data Catalog
Infer Schemas
4
Connects to data store
3
Connection
Crawler
1
Runs
Custom Classifiers
2
Built-in Classifiers
Writes Metadata
5
Data Catalog
Database
Database
Tables (Metadata)
Tables (Metadata)

The following is the general workflow for how a crawler populates the AWS Glue Data Catalog:

1. A crawler runs any custom *classifiers* that you choose to infer the format and schema of your data. You provide the code for custom classifiers, and they run in the order that you specify.

   The first custom classifier to successfully recognize the structure of your data is used to create a schema. Custom classifiers lower in the list are skipped.
2. If no custom classifier matches your data's schema, built-in classifiers try to recognize your data's schema. An example of a built-in classifier is one that recognizes JSON.
3. The crawler connects to the data store. Some data stores require connection properties for crawler access.
4. The inferred schema is created for your data.

5. The crawler writes metadata to the Data Catalog. A table definition contains metadata about the data in your data store. The table is written to a database, which is a container of tables in the Data Catalog. Attributes of a table include classification, which is a label created by the classifier that inferred the table schema.

Topics

- How crawlers work
- How does a crawler determine when to create partitions?
- Supported data sources for crawling
- Crawler prerequisites
- Defining and managing classifiers
- Configuring a crawler
- Scheduling a crawler
- Viewing crawler results and details
- Customizing crawler behavior
- Tutorial: Adding an AWS Glue crawler

How crawlers work

When a crawler runs, it takes the following actions to interrogate a data store:

- **Classifies data to determine the format, schema, and associated properties of the raw data** – You can configure the results of classification by creating a custom classifier.
- **Groups data into tables or partitions** – Data is grouped based on crawler heuristics.
- **Writes metadata to the Data Catalog** – You can configure how the crawler adds, updates, and deletes tables and partitions.

When you define a crawler, you choose one or more classifiers that evaluate the format of your data to infer a schema. When the crawler runs, the first classifier in your list to successfully recognize your data store is used to create a schema for your table. You can use built-in classifiers or define your own. You define your custom classifiers in a separate operation, before you define the crawlers. AWS Glue provides built-in classifiers to infer schemas from common files with formats that include JSON, CSV, and Apache Avro. For the current list of built-in classifiers in AWS Glue, see Built-in classifiers.

The metadata tables that a crawler creates are contained in a database when you define a crawler. If your crawler does not specify a database, your tables are placed in the default database. In addition, each table has a classification column that is filled in by the classifier that first successfully recognized the data store.

If the file that is crawled is compressed, the crawler must download it to process it. When a crawler runs, it interrogates files to determine their format and compression type and writes these properties into the Data Catalog. Some file formats (for example, Apache Parquet) enable you to compress parts of the file as it is written. For these files, the compressed data is an internal component of the file, and AWS Glue does not populate the `compressionType` property when it writes tables into the Data Catalog. In contrast, if an *entire file* is compressed by a compression algorithm (for example, gzip), then the `compressionType` property is populated when tables are written into the Data Catalog.

The crawler generates the names for the tables that it creates. The names of the tables that are stored in the AWS Glue Data Catalog follow these rules:

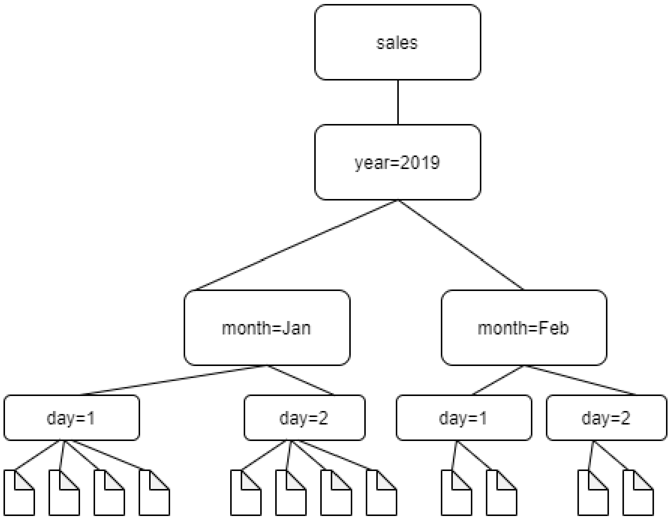
- Only alphanumeric characters and underscore (_) are allowed.
- Any custom prefix cannot be longer than 64 characters.
- The maximum length of the name cannot be longer than 128 characters. The crawler truncates generated names to fit within the limit.
- If duplicate table names are encountered, the crawler adds a hash string suffix to the name.

If your crawler runs more than once, perhaps on a schedule, it looks for new or changed files or tables in your data store. The output of the crawler includes new tables and partitions found since a previous run.

How does a crawler determine when to create partitions?

When an AWS Glue crawler scans Amazon S3 data stpre and detects multiple folders in a bucket, it determines the root of a table in the folder structure and which folders are partitions of a table. The name of the table is based on the Amazon S3 prefix or folder name. You provide an **Include path** that points to the folder level to crawl. When the majority of schemas at a folder level are similar, the crawler creates partitions of a table instead of separate tables. To influence the crawler to create separate tables, add each table's root folder as a separate data store when you define the crawler.

For example, consider the following Amazon S3 folder structure.



The paths to the four lowest level folders are the following:

```
S3://sales/year=2019/month=Jan/day=1
S3://sales/year=2019/month=Jan/day=2
S3://sales/year=2019/month=Feb/day=1
S3://sales/year=2019/month=Feb/day=2
```

Assume that the crawler target is set at `Sales`, and that all files in the `day=`*n* folders have the same format (for example, JSON, not encrypted), and have the same or very similar schemas. The crawler will create a single table with four partitions, with partition keys `year`, `month`, and `day`.

In the next example, consider the following Amazon S3 structure:

```

s3://bucket01/folder1/table1/partition1/file.txt
s3://bucket01/folder1/table1/partition2/file.txt
s3://bucket01/folder1/table1/partition3/file.txt
s3://bucket01/folder1/table2/partition4/file.txt
s3://bucket01/folder1/table2/partition5/file.txt
```

If the schemas for files under `table1` and `table2` are similar, and a single data store is defined in the crawler with **Include path** `s3://bucket01/folder1/`, the crawler creates a single table with two partition key columns. The first partition key column contains `table1` and `table2`, and the second partition key column contains `partition1` through `partition3` for the `table1` partition and `partition4` and `partition5` for the `table2` partition. To create two separate tables, define the crawler with two data stores. In this example, define the first **Include path** as `s3://bucket01/folder1/table1/` and the second as `s3://bucket01/folder1/table2`.

Note

In Amazon Athena, each table corresponds to an Amazon S3 prefix with all the objects in it. If objects have different schemas, Athena does not recognize different objects within the same prefix as separate tables. This can happen if a crawler creates multiple tables from the same Amazon S3 prefix. This might lead to queries in Athena that return zero results. For Athena to properly recognize and query tables, create the crawler with a separate **Include path** for each different table schema in the Amazon S3 folder structure. For more information, see Best Practices When Using Athena with AWS Glue and this AWS Knowledge Center article.

Supported data sources for crawling

Crawlers can crawl the following file-based and table-based data stores.

Access type that crawler uses	Data stores
Native client	• Amazon Simple Storage Service (Amazon S3) • Amazon DynamoDB • Delta Lake 2.0.x • Apache Iceberg 1.5 • Apache Hudi 0.14
JDBC	Amazon Redshift Snowflake Within Amazon Relational Database Service (Amazon RDS) or external to Amazon RDS: • Amazon Aurora • MariaDB • Microsoft SQL Server • MySQL • Oracle

Access type that crawler uses	Data stores
	- PostgreSQL
MongoDB client	- MongoDB - MongoDB Atlas - Amazon DocumentDB (with MongoDB compatibility)

Note

Currently AWS Glue does not support crawlers for data streams.

For JDBC, MongoDB, MongoDB Atlas, and Amazon DocumentDB (with MongoDB compatibility) data stores, you must specify an AWS Glue *connection* that the crawler can use to connect to the data store. For Amazon S3, you can optionally specify a connection of type Network. A connection is a Data Catalog object that stores connection information, such as credentials, URL, Amazon Virtual Private Cloud information, and more. For more information, see Connecting to data.

The following are the versions of drivers supported by the crawler:

Product	Crawler supported driver
PostgreSQL	42.2.1
Amazon Aurora	Same as native crawler drivers
MariaDB	8.0.13
Microsoft SQL Server	6.1.0
MySQL	8.0.13
Oracle	11.2.2
Amazon Redshift	4.1
Snowflake	3.13.20

Product	Crawler supported driver
MongoDB	4.7.2
MongoDB Atlas	4.7.2

The following are notes about the various data stores.

Amazon S3

You can choose to crawl a path in your account or in another account. If all the Amazon S3 files in a folder have the same schema, the crawler creates one table. Also, if the Amazon S3 object is partitioned, only one metadata table is created and partition information is added to the Data Catalog for that table.

Amazon S3 and Amazon DynamoDB

Crawlers use an AWS Identity and Access Management (IAM) role for permission to access your data stores. *The role you pass to the crawler must have permission to access Amazon S3 paths and Amazon DynamoDB tables that are crawled*.

Amazon DynamoDB

When defining a crawler using the AWS Glue console, you specify one DynamoDB table. If you're using the AWS Glue API, you can specify a list of tables. You can choose to crawl only a small sample of the data to reduce crawler run times.

Delta Lake

For each Delta Lake data store, you specify how to create the Delta tables:

- **Create Native tables**: Allow integration with query engines that support querying of the Delta transaction log directly. For more information, see Querying Delta Lake tables.
- **Create Symlink tables**: Create a `_symlink_manifest` folder with manifest files partitioned by the partition keys, based on the specified configuration parameters.

Iceberg

For each Iceberg data store, you specify an Amazon S3 path that contains the metadata for your Iceberg tables. If crawler discovers Iceberg table metadata, it registers it in the Data Catalog. You can set a schedule for the crawler to keep the tables updated.

You can define these parameters for the data store:

- **Exclusions**: Allows you to skip certain folders.
- **Maximum Traversal Depth**: Sets the depth limit the crawler can crawl in your Amazon S3 bucket. The default maximum traversal depth is 10 and the maximum depth you can set is 20.

Hudi

For each Hudi data store, you specify an Amazon S3 path that contains the metadata for your Hudi tables. If crawler discovers Hudi table metadata, it registers it in the Data Catalog. You can set a schedule for the crawler to keep the tables updated.

You can define these parameters for the data store:

- **Exclusions**: Allows you to skip certain folders.
- **Maximum Traversal Depth**: Sets the depth limit the crawler can crawl in your Amazon S3 bucket. The default maximum traversal depth is 10 and the maximum depth you can set is 20.

> **Note**
>
> Timestamp columns with `millis` as logical types will be interpreted as `bigint`, due to an incompatibility with Hudi 0.13.1 and timestamp types. A resolution may be provided in the upcoming Hudi release.

Hudi tables are categorized as follows, with specific implications for each:

- Copy on Write (CoW): Data is stored in a columnar format (Parquet), and each update creates a new version of files during a write.
- Merge on Read (MoR): Data is stored using a combination of columnar (Parquet) and row-based (Avro) formats. Updates are logged to row-based delta files and are compacted as needed to create new versions of the columnar files.

With CoW datasets, each time there is an update to a record, the file that contains the record is rewritten with the updated values. With a MoR dataset, each time there is an update, Hudi writes only the row for the changed record. MoR is better suited for write- or change-heavy workloads with fewer reads. CoW is better suited for read-heavy workloads on data that change less frequently.

Hudi provides three query types for accessing the data:

- Snapshot queries: Queries that see the latest snapshot of the table as of a given commit or compaction action. For MoR tables, snapshot queries expose the most recent state of the table by merging the base and delta files of the latest file slice at the time of the query.

- Incremental queries: Queries only see new data written to the table, since a given commit/compaction. This effectively provides change streams to enable incremental data pipelines.
- Read optimized queries: For MoR tables, queries see the latest data compacted. For CoW tables, queries see the latest data committed.

For Copy-On-Write tables, the crawlers creates a single table in the Data Catalog with the ReadOptimized serde `org.apache.hudi.hadoop.HoodieParquetInputFormat`.

For Merge-On-Read tables, the crawler creates two tables in the Data Catalog for the same table location:

- A table with suffix `_ro` which uses the ReadOptimized serde `org.apache.hudi.hadoop.HoodieParquetInputFormat`.
- A table with suffix `_rt` which uses the RealTime Serde allowing for Snapshot queries: `org.apache.hudi.hadoop.realtime.HoodieParquetRealtimeInputFormat`.

MongoDB and Amazon DocumentDB (with MongoDB compatibility)

MongoDB versions 3.2 and later are supported. You can choose to crawl only a small sample of the data to reduce crawler run times.

Relational database

Authentication is with a database user name and password. Depending on the type of database engine, you can choose which objects are crawled, such as databases, schemas, and tables.

Snowflake

The Snowflake JDBC crawler supports crawling the Table, External Table, View, and Materialized View. The Materialized View Definition will not be populated.

For Snowflake external tables, the crawler only will crawl if it points to an Amazon S3 location. In addition to the the table schema, the crawler will also crawl the Amazon S3 location, file format and output as table parameters in the Data Catalog table. Note that the partition information of the partitioned external table is not populated.

ETL is currently not supported for Data Catalog tables created using the Snowflake crawler.

Crawler prerequisites

The crawler assumes the permissions of the AWS Identity and Access Management (IAM) role that you specify when you define it. This IAM role must have permissions to extract data from your data store and write to the Data Catalog. The AWS Glue console lists only IAM roles that have attached a trust policy for the AWS Glue principal service. From the console, you can also create an IAM role with an IAM policy to access Amazon S3 data stores accessed by the crawler. For more information about providing roles for AWS Glue, see Identity-based policies for AWS Glue.

> **Note**
>
> When crawling a Delta Lake data store, you must have Read/Write permissions to the Amazon S3 location.

For your crawler, you can create a role and attach the following policies:

- The `AWSGlueServiceRole` AWS managed policy, which grants the required permissions on the Data Catalog
- An inline policy that grants permissions on the data source.
- An inline policy that grants `iam:PassRole` permission on the role.

A quicker approach is to let the AWS Glue console crawler wizard create a role for you. The role that it creates is specifically for the crawler, and includes the `AWSGlueServiceRole` AWS managed policy plus the required inline policy for the specified data source.

If you specify an existing role for a crawler, ensure that it includes the `AWSGlueServiceRole` policy or equivalent (or a scoped down version of this policy), plus the required inline policies. For example, for an Amazon S3 data store, the inline policy would at a minimum be the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "s3:GetObject"
      ],
      "Resource": [
        "arn:aws:s3:::bucket/object*"
      ]
    }
  ]
}
```

For an Amazon DynamoDB data store, the policy would at a minimum be the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "dynamodb:DescribeTable",
        "dynamodb:Scan"
      ],
      "Resource": [
        "arn:aws:dynamodb:us-east-1:111122223333:table/table-name*"
      ]
    }
  ]
}
```

In addition, if the crawler reads AWS Key Management Service (AWS KMS) encrypted Amazon S3 data, then the IAM role must have decrypt permission on the AWS KMS key. For more information, see Step 2: Create an IAM role for AWS Glue.

Defining and managing classifiers

A classifier reads the data in a data store. If it recognizes the format of the data, it generates a schema. The classifier also returns a certainty number to indicate how certain the format recognition was.

AWS Glue provides a set of built-in classifiers, but you can also create custom classifiers. AWS Glue invokes custom classifiers first, in the order that you specify in your crawler definition. Depending on the results that are returned from custom classifiers, AWS Glue might also invoke built-in classifiers. If a classifier returns `certainty=1.0` during processing, it indicates that it's

100 percent certain that it can create the correct schema. AWS Glue then uses the output of that classifier.

If no classifier returns `certainty=1.0`, AWS Glue uses the output of the classifier that has the highest certainty. If no classifier returns a certainty greater than `0.0`, AWS Glue returns the default classification string of `UNKNOWN`.

When do I use a classifier?

You use classifiers when you crawl a data store to define metadata tables in the AWS Glue Data Catalog. You can set up your crawler with an ordered set of classifiers. When the crawler invokes a classifier, the classifier determines whether the data is recognized. If the classifier can't recognize the data or is not 100 percent certain, the crawler invokes the next classifier in the list to determine whether it can recognize the data.

For more information about creating a classifier using the AWS Glue console, see Creating classifiers using the AWS Glue console.

Custom classifiers

The output of a classifier includes a string that indicates the file's classification or format (for example, `json`) and the schema of the file. For custom classifiers, you define the logic for creating the schema based on the type of classifier. Classifier types include defining schemas based on grok patterns, XML tags, and JSON paths.

If you change a classifier definition, any data that was previously crawled using the classifier is not reclassified. A crawler keeps track of previously crawled data. New data is classified with the updated classifier, which might result in an updated schema. If the schema of your data has evolved, update the classifier to account for any schema changes when your crawler runs. To reclassify data to correct an incorrect classifier, create a new crawler with the updated classifier.

For more information about creating custom classifiers in AWS Glue, see Writing custom classifiers for diverse data formats.

> **Note**
>
> If your data format is recognized by one of the built-in classifiers, you don't need to create a custom classifier.

Built-in classifiers

AWS Glue provides built-in classifiers for various formats, including JSON, CSV, web logs, and many database systems.

If AWS Glue doesn't find a custom classifier that fits the input data format with 100 percent certainty, it invokes the built-in classifiers in the order shown in the following table. The built-in classifiers return a result to indicate whether the format matches (`certainty=1.0`) or does not match (`certainty=0.0`). The first classifier that has `certainty=1.0` provides the classification string and schema for a metadata table in your Data Catalog.

Classifier type	Classification string	Notes
Apache Avro	`avro`	Reads the schema at the beginning of the file to determine format.
Apache ORC	`orc`	Reads the file metadata to determine format.
Apache Parquet	`parquet`	Reads the schema at the end of the file to determine format.
JSON	`json`	Reads the beginning of the file to determine format.
Binary JSON	`bson`	Reads the beginning of the file to determine format.
XML	`xml`	Reads the beginning of the file to determine format. AWS Glue determines the table schema based on XML tags in the document. For information about creating a custom XML classifier to specify rows in the document, see Writing XML custom classifiers.
Amazon Ion	`ion`	Reads the beginning of the file to determine format.

Classifier type	Classification string	Notes
Combined Apache log	`combined_apache`	Determines log formats through a grok pattern.
Apache log	`apache`	Determines log formats through a grok pattern.
Linux kernel log	`linux_kernel`	Determines log formats through a grok pattern.
Microsoft log	`microsoft_log`	Determines log formats through a grok pattern.
Ruby log	`ruby_logger`	Reads the beginning of the file to determine format.
Squid 3.x log	`squid`	Reads the beginning of the file to determine format.
Redis monitor log	`redismonlog`	Reads the beginning of the file to determine format.
Redis log	`redislog`	Reads the beginning of the file to determine format.
CSV	`csv`	Checks for the following delimiters: comma (,), pipe (|), tab (\t), semicolon (;), and Ctrl-A (\u0001). Ctrl-A is the Unicode control character for `Start Of Heading`.
Amazon Redshift	`redshift`	Uses JDBC connection to import metadata.
MySQL	`mysql`	Uses JDBC connection to import metadata.
PostgreSQL	`postgresql`	Uses JDBC connection to import metadata.
Oracle database	`oracle`	Uses JDBC connection to import metadata.
Microsoft SQL Server	`sqlserver`	Uses JDBC connection to import metadata.

Classifier type	Classification string	Notes
Amazon DynamoDB	`dynamodb`	Reads data from the DynamoDB table.

Files in the following compressed formats can be classified:

- ZIP (supported for archives containing only a single file). Note that Zip is not well-supported in other services (because of the archive).
- BZIP
- GZIP
- LZ4
- Snappy (supported for both standard and Hadoop native Snappy formats)

Built-in CSV classifier

The built-in CSV classifier parses CSV file contents to determine the schema for an AWS Glue table. This classifier checks for the following delimiters:

- Comma (,)
- Pipe (|)
- Tab (\t)
- Semicolon (;)
- Ctrl-A (\u0001)

  Ctrl-A is the Unicode control character for `Start Of Heading`.

To be classified as CSV, the table schema must have at least two columns and two rows of data. The CSV classifier uses a number of heuristics to determine whether a header is present in a given file. If the classifier can't determine a header from the first row of data, column headers are displayed as `col1`, `col2`, `col3`, and so on. The built-in CSV classifier determines whether to infer a header by evaluating the following characteristics of the file:

- Every column in a potential header parses as a STRING data type.

- Except for the last column, every column in a potential header has content that is fewer than 150 characters. To allow for a trailing delimiter, the last column can be empty throughout the file.
- Every column in a potential header must meet the AWS Glue `regex` requirements for a column name.
- The header row must be sufficiently different from the data rows. To determine this, one or more of the rows must parse as other than STRING type. If all columns are of type STRING, then the first row of data is not sufficiently different from subsequent rows to be used as the header.

> **Note**
>
> If the built-in CSV classifier does not create your AWS Glue table as you want, you might be able to use one of the following alternatives:
>
> - Change the column names in the Data Catalog, set the `SchemaChangePolicy` to LOG, and set the partition output configuration to `InheritFromTable` for future crawler runs.
> - Create a custom grok classifier to parse the data and assign the columns that you want.
> - The built-in CSV classifier creates tables referencing the `LazySimpleSerDe` as the serialization library, which is a good choice for type inference. However, if the CSV data contains quoted strings, edit the table definition and change the SerDe library to `OpenCSVSerDe`. Adjust any inferred types to STRING, set the `SchemaChangePolicy` to LOG, and set the partitions output configuration to `InheritFromTable` for future crawler runs. For more information about SerDe libraries, see SerDe Reference in the Amazon Athena User Guide.

Writing custom classifiers for diverse data formats

You can provide a custom classifier to classify your data in AWS Glue. You can create a custom classifier using a grok pattern, an XML tag, JavaScript Object Notation (JSON), or comma-separated values (CSV). An AWS Glue crawler calls a custom classifier. If the classifier recognizes the data, it returns the classification and schema of the data to the crawler. You might need to define a custom classifier if your data doesn't match any built-in classifiers, or if you want to customize the tables that are created by the crawler.

For more information about creating a classifier using the AWS Glue console, see Creating classifiers using the AWS Glue console.

AWS Glue runs custom classifiers before built-in classifiers, in the order you specify. When a crawler finds a classifier that matches the data, the classification string and schema are used in the definition of tables that are written to your AWS Glue Data Catalog.

Topics

- Writing grok custom classifiers
- Writing XML custom classifiers
- Writing JSON custom classifiers
- Writing CSV custom classifiers

Writing grok custom classifiers

Grok is a tool that is used to parse textual data given a matching pattern. A grok pattern is a named set of regular expressions (regex) that are used to match data one line at a time. AWS Glue uses grok patterns to infer the schema of your data. When a grok pattern matches your data, AWS Glue uses the pattern to determine the structure of your data and map it into fields.

AWS Glue provides many built-in patterns, or you can define your own. You can create a grok pattern using built-in patterns and custom patterns in your custom classifier definition. You can tailor a grok pattern to classify custom text file formats.

> **Note**
>
> AWS Glue grok custom classifiers use the `GrokSerDe` serialization library for tables created in the AWS Glue Data Catalog. If you are using the AWS Glue Data Catalog with Amazon Athena, Amazon EMR, or Redshift Spectrum, check the documentation about those services for information about support of the `GrokSerDe`. Currently, you might encounter problems querying tables created with the `GrokSerDe` from Amazon EMR and Redshift Spectrum.

The following is the basic syntax for the components of a grok pattern:

```
%{PATTERN:field-name}
```

Data that matches the named PATTERN is mapped to the `field-name` column in the schema, with a default data type of `string`. Optionally, the data type for the field can be cast to `byte`, `boolean`, `double`, `short`, `int`, `long`, or `float` in the resulting schema.

```
%{PATTERN:field-name:data-type}
```

For example, to cast a num field to an `int` data type, you can use this pattern:

```
%{NUMBER:num:int}
```

Patterns can be composed of other patterns. For example, you can have a pattern for a SYSLOG timestamp that is defined by patterns for month, day of the month, and time (for example, `Feb 1 06:25:43`). For this data, you might define the following pattern:

```
SYSLOGTIMESTAMP %{MONTH} +%{MONTHDAY} %{TIME}
```

> **Note**
>
> Grok patterns can process only one line at a time. Multiple-line patterns are not supported. Also, line breaks within a pattern are not supported.

Custom values for grok classifier

When you define a grok classifier, you supply the following values to create the custom classifier.

Name

Name of the classifier.

Classification

The text string that is written to describe the format of the data that is classified; for example, `special-logs`.

Grok pattern

The set of patterns that are applied to the data store to determine whether there is a match. These patterns are from AWS Glue built-in patterns and any custom patterns that you define.

The following is an example of a grok pattern:

```
%{TIMESTAMP_ISO8601:timestamp} \[%{MESSAGEPREFIX:message_prefix}\]
 %{CRAWLERLOGLEVEL:loglevel} : %{GREEDYDATA:message}
```

When the data matches `TIMESTAMP_ISO8601`, a schema column `timestamp` is created. The behavior is similar for the other named patterns in the example.

Custom patterns

Optional custom patterns that you define. These patterns are referenced by the grok pattern that classifies your data. You can reference these custom patterns in the grok pattern that is applied to your data. Each custom component pattern must be on a separate line. Regular expression (regex) syntax is used to define the pattern.

The following is an example of using custom patterns:

```
CRAWLERLOGLEVEL (BENCHMARK|ERROR|WARN|INFO|TRACE)
MESSAGEPREFIX .*-.*-.*-.*-.*
```

The first custom named pattern, `CRAWLERLOGLEVEL`, is a match when the data matches one of the enumerated strings. The second custom pattern, `MESSAGEPREFIX`, tries to match a message prefix string.

AWS Glue keeps track of the creation time, last update time, and version of your classifier.

Built-in patterns

AWS Glue provides many common patterns that you can use to build a custom classifier. You add a named pattern to the `grok pattern` in a classifier definition.

The following list consists of a line for each pattern. In each line, the pattern name is followed its definition. Regular expression (regex) syntax is used in defining the pattern.

```
#<noloc>&GLU;</noloc> Built-in patterns
 USERNAME [a-zA-Z0-9._-]+
 USER %{USERNAME:UNWANTED}
 INT (?:[+-]?(?:[0-9]+))
 BASE10NUM (?<![0-9.+-])(?>[+-]?(?:(?:[0-9]+(?:\.[0-9]+)?)|(?:\.[0-9]+)))
 NUMBER (?:%{BASE10NUM:UNWANTED})
 BASE16NUM (?<![0-9A-Fa-f])(?:[+-]?(?:0x)?(?:[0-9A-Fa-f]+))
 BASE16FLOAT \b(?<![0-9A-Fa-f.])(?:[+-]?(?:0x)?(?:(?:[0-9A-Fa-f]+(?:\.[0-9A-Fa-f]*)?)|(?:\.[0-9A-Fa-f]+)))\b
 BOOLEAN (?i)(true|false)

 POSINT \b(?:[1-9][0-9]*)\b
 NONNEGINT \b(?:[0-9]+)\b
 WORD \b\w+\b
 NOTSPACE \S+
 SPACE \s*
 DATA .*?
 GREEDYDATA .*
 #QUOTEDSTRING (?:(?<!\\)(?:"(?:\\.|[^\\"])*"|(?:'(?:\\.|[^\\'])*')|(?:`(?:\\.|[^\\`])*`)))
 QUOTEDSTRING (?>(?<!\\)(?>"(?>\\.|[^\\"]+)+"|""|(?>'(?>\\.|[^\\']+)+')|''|(?>`(?>\\.|[^\\`]+)+`)|``))
 UUID [A-Fa-f0-9]{8}-(?:[A-Fa-f0-9]{4}-){3}[A-Fa-f0-9]{12}

 # Networking
 MAC (?:%{CISCOMAC:UNWANTED}|%{WINDOWSMAC:UNWANTED}|%{COMMONMAC:UNWANTED})
 CISCOMAC (?:(?:[A-Fa-f0-9]{4}\.){2}[A-Fa-f0-9]{4})
 WINDOWSMAC (?:(?:[A-Fa-f0-9]{2}-){5}[A-Fa-f0-9]{2})
 COMMONMAC (?:(?:[A-Fa-f0-9]{2}:){5}[A-Fa-f0-9]{2})
 IPV6 ((([0-9A-Fa-f]{1,4}:){7}([0-9A-Fa-f]{1,4}|:))|(([0-9A-Fa-f]{1,4}:){6}(:[0-9A-Fa-f]{1,4}|((25[0-5]|2[0-4]\d|1\d\d|[1-9]?\d)(\.(25[0-5]|2[0-4]\d|1\d\d|[1-9]?\d)){3})|:))|(([0-9A-Fa-f]{1,4}:){5}(((:[0-9A-Fa-f]{1,4}){1,2})|:((25[0-5]|2[0-4]\d|1\d\d|[1-9]?\d)(\.(25[0-5]|2[0-4]\d|1\d\d|[1-9]?\d)){3})|:))|(([0-9A-Fa-f]{1,4}:){4}(((:[0-9A-Fa-f]{1,4}){1,3})|((:[0-9A-Fa-f]{1,4})?:((25[0-5]|2[0-4]\d|1\d\d|[1-9]?\d)(\.(25[0-5]|2[0-4]\d|1\d\d|[1-9]?\d)){3}))|:))|(([0-9A-Fa-f]{1,4}:){3}(((:[0-9A-Fa-f]{1,4}){1,4})|((:[0-9A-Fa-f]{1,4}){0,2}:((25[0-5]|2[0-4]\d|1\d\d|[1-9]?\d)(\.(25[0-5]|2[0-4]\d|1\d\d|[1-9]?\d)){3}))|:))|(([0-9A-Fa-f]{1,4}:){2}(((:[0-9A-Fa-f]{1,4}){1,5})|((:[0-9A-Fa-f]{1,4}){0,3}:((25[0-5]|2[0-4]\d|1\d\d|[1-9]?\d)(\.(25[0-5]|2[0-4]\d|1\d\d|[1-9]?\d)){3}))|:))|(([0-9A-Fa-f]{1,4}:){1}(((:[0-9A-Fa-f]{1,4}){1,6})|((:[0-9A-Fa-f]{1,4}){0,4}:((25[0-5]|2[0-4]\d|1\d\d|[1-9]?\d)(\.(25[0-5]|2[0-4]\d|1\d\d|[1-9]?\d)){3}))|:))|(:(((:[0-9A-Fa-f]{1,4}){1,7})|((:[0-9A-Fa-f]{1,4}){0,5}:((25[0-5]|2[0-4]\d|1\d\d|[1-9]?\d)(\.(25[0-5]|2[0-4]\d|1\d\d|[1-9]?\d)){3}))|:)))(%.+)?
 IPV4 (?<![0-9])(?:(?:25[0-5]|2[0-4][0-9]|[0-1]?[0-9]{1,2})[.](?:25[0-5]|2[0-4][0-9]|[0-1]?[0-9]{1,2})[.](?:25[0-5]|2[0-4][0-9]|[0-1]?[0-9]{1,2})[.](?:25[0-5]|2[0-4][0-9]|[0-1]?[0-9]{1,2}))(?![0-9])
 IP (?:%{IPV6:UNWANTED}|%{IPV4:UNWANTED})
 HOSTNAME \b(?:[0-9A-Za-z][0-9A-Za-z-_]{0,62})(?:\.(?:[0-9A-Za-z][0-9A-Za-z-_]{0,62}))*(\.?|\b)
 HOST %{HOSTNAME:UNWANTED}
 IPORHOST (?:%{HOSTNAME:UNWANTED}|%{IP:UNWANTED})
 HOSTPORT (?:%{IPORHOST}:%{POSINT:PORT})
 # paths
 PATH (?:%{UNIXPATH}|%{WINPATH})
 UNIXPATH (?>/(?>[\w_%!$@:.,~-]+|\\.)*)+
 #UNIXPATH (?<![\w\/])(?:/[^\/\s?*]*)+
 TTY (?:/dev/(pts|tty([pq])?)(\w+)?/?(?:[0-9]+))
 WINPATH (?>[A-Za-z]+:|\\)(?:\\[^\\?*]*)+
 URIPROTO [A-Za-z]+(\+[A-Za-z+]+)?
 URIHOST %{IPORHOST}(?::%{POSINT:port})?
 # uripath comes loosely from RFC1738, but mostly from what Firefox
 # doesn't turn into %XX
 URIPATH (?:/[A-Za-z0-9$.+!*'(){},~:;=@#%_\-]*)+
 #URIPARAM \?(?:[A-Za-z0-9]+(?:=(?:[^&]*))?(?:&(?:[A-Za-z0-9]+(?:=(?:[^&]*))?)?)*)?
 URIPARAM \?[A-Za-z0-9$.+!*'|(){},~@#%&/=:;_?\-\[\]]*
 URIPATHPARAM %{URIPATH}(?:%{URIPARAM})?
 URI %{URIPROTO}://(?:%{USER}(?::[^@]*)?@)?(?:%{URIHOST})?(?:%{URIPATHPARAM})?

 # Months: January, Feb, 3, 03, 12, December
 MONTH \b(?:Jan(?:uary)?|Feb(?:ruary)?|Mar(?:ch)?|Apr(?:il)?|May|Jun(?:e)?|Jul(?:y)?|
Aug(?:ust)?|Sep(?:tember)?|Oct(?:ober)?|Nov(?:ember)?|Dec(?:ember)?)\b
 MONTHNUM (?:0?[1-9]|1[0-2])
 MONTHNUM2 (?:0[1-9]|1[0-2])
 MONTHDAY (?:(?:0[1-9])|(?:[12][0-9])|(?:3[01])|[1-9])

 # Days: Monday, Tue, Thu, etc...
 DAY (?:Mon(?:day)?|Tue(?:sday)?|Wed(?:nesday)?|Thu(?:rsday)?|Fri(?:day)?|
Sat(?:urday)?|Sun(?:day)?)

 # Years?
 YEAR (?>\d\d){1,2}
 # Time: HH:MM:SS
 #TIME \d{2}:\d{2}(?::\d{2}(?:\.\d+)?)?
 # TIME %{POSINT<24}:%{POSINT<60}(?::%{POSINT<60}(?:\.%{POSINT})?)?
 HOUR (?:2[0123]|[01]?[0-9])
 MINUTE (?:[0-5][0-9])
 # '60' is a leap second in most time standards and thus is valid.
 SECOND (?:(?:[0-5]?[0-9]|60)(?:[:.,][0-9]+)?)
 TIME (?!<[0-9])%{HOUR}:%{MINUTE}(?::%{SECOND})(?![0-9])
 # datestamp is YYYY/MM/DD-HH:MM:SS.UUUU (or something like it)
 DATE_US %{MONTHNUM}[/-]%{MONTHDAY}[/-]%{YEAR}
 DATE_EU %{MONTHDAY}[./-]%{MONTHNUM}[./-]%{YEAR}
 DATESTAMP_US %{DATE_US}[- ]%{TIME}
 DATESTAMP_EU %{DATE_EU}[- ]%{TIME}
 ISO8601_TIMEZONE (?:Z|[+-]%{HOUR}(?::?%{MINUTE}))
 ISO8601_SECOND (?:%{SECOND}|60)
 TIMESTAMP_ISO8601 %{YEAR}-%{MONTHNUM}-%{MONTHDAY}[T ]%{HOUR}:?%{MINUTE}(?::?
%{SECOND})?%{ISO8601_TIMEZONE}?
 TZ (?:[PMCE][SD]T|UTC)
 DATESTAMP_RFC822 %{DAY} %{MONTH} %{MONTHDAY} %{YEAR} %{TIME} %{TZ}
 DATESTAMP_RFC2822 %{DAY}, %{MONTHDAY} %{MONTH} %{YEAR} %{TIME} %{ISO8601_TIMEZONE}
 DATESTAMP_OTHER %{DAY} %{MONTH} %{MONTHDAY} %{TIME} %{TZ} %{YEAR}
 DATESTAMP_EVENTLOG %{YEAR}%{MONTHNUM2}%{MONTHDAY}%{HOUR}%{MINUTE}%{SECOND}
 CISCOTIMESTAMP %{MONTH} %{MONTHDAY} %{TIME}

 # Syslog Dates: Month Day HH:MM:SS
 SYSLOGTIMESTAMP %{MONTH} +%{MONTHDAY} %{TIME}
 PROG (?:[\w._/%-]+)
 SYSLOGPROG %{PROG:program}(?:\[%{POSINT:pid}\])?
 SYSLOGHOST %{IPORHOST}
 SYSLOGFACILITY <%{NONNEGINT:facility}.%{NONNEGINT:priority}>
 HTTPDATE %{MONTHDAY}/%{MONTH}/%{YEAR}:%{TIME} %{INT}

 # Shortcuts
 QS %{QUOTEDSTRING:UNWANTED}

 # Log formats
 SYSLOGBASE %{SYSLOGTIMESTAMP:timestamp} (?:%{SYSLOGFACILITY} )?%{SYSLOGHOST:logsource}
 %{SYSLOGPROG}:

 MESSAGESLOG %{SYSLOGBASE} %{DATA}

 COMMONAPACHELOG %{IPORHOST:clientip} %{USER:ident} %{USER:auth}
 \[%{HTTPDATE:timestamp}\] "(?:%{WORD:verb} %{NOTSPACE:request}(?: HTTP/
%{NUMBER:httpversion})?|%{DATA:rawrequest})" %{NUMBER:response} (?:%{Bytes:bytes=
%{NUMBER}|-})
 COMBINEDAPACHELOG %{COMMONAPACHELOG} %{QS:referrer} %{QS:agent}
 COMMONAPACHELOG_DATATYPED %{IPORHOST:clientip} %{USER:ident;boolean} %{USER:auth}
 \[%{HTTPDATE:timestamp;date;dd/MMM/yyyy:HH:mm:ss Z}\] "(?:%{WORD:verb;string}
 %{NOTSPACE:request}(?: HTTP/%{NUMBER:httpversion;float})?|%{DATA:rawrequest})"
 %{NUMBER:response;int} (?:%{NUMBER:bytes;long}|-)


 # Log Levels
 LOGLEVEL ([A|a]lert|ALERT|[T|t]race|TRACE|[D|d]ebug|DEBUG|[N|n]otice|NOTICE|[I|i]nfo|
INFO|[W|w]arn?(?:ing)?|WARN?(?:ING)?|[E|e]rr?(?:or)?|ERR?(?:OR)?|[C|c]rit?(?:ical)?|
CRIT?(?:ICAL)?|[F|f]atal|FATAL|[S|s]evere|SEVERE|EMERG(?:ENCY)?|[Ee]merg(?:ency)?)
```

Writing XML custom classifiers

XML defines the structure of a document with the use of tags in the file. With an XML custom classifier, you can specify the tag name used to define a row.

Custom classifier values for an XML classifier

When you define an XML classifier, you supply the following values to AWS Glue to create the classifier. The classification field of this classifier is set to `xml`.

Name

Name of the classifier.

Row tag

The XML tag name that defines a table row in the XML document, without angle brackets < >. The name must comply with XML rules for a tag.

> **Note**
>
> The element containing the row data **cannot** be a self-closing empty element. For example, this empty element is **not** parsed by AWS Glue:
>
> ```
>             <row att1=”xx” att2=”yy” />
> ```
>
> Empty elements can be written as follows:
>
> ```
>             <row att1=”xx” att2=”yy”> </row>
> ```

AWS Glue keeps track of the creation time, last update time, and version of your classifier.

For example, suppose that you have the following XML file. To create an AWS Glue table that only contains columns for author and title, create a classifier in the AWS Glue console with **Row tag** as `AnyCompany`. Then add and run a crawler that uses this custom classifier.

```
<?xml version="1.0"?>
<catalog>
   <book id="bk101">
     <AnyCompany>
       <author>Rivera, Martha</author>
       <title>AnyCompany Developer Guide</title>
     </AnyCompany>
   </book>
   <book id="bk102">
     <AnyCompany>
       <author>Stiles, John</author>
       <title>Style Guide for AnyCompany</title>
     </AnyCompany>
   </book>
</catalog>
```

Writing JSON custom classifiers

JSON is a data-interchange format. It defines data structures with name-value pairs or an ordered list of values. With a JSON custom classifier, you can specify the JSON path to a data structure that is used to define the schema for your table.

Custom classifier values in AWS Glue

When you define a JSON classifier, you supply the following values to AWS Glue to create the classifier. The classification field of this classifier is set to `json`.

Name

Name of the classifier.

JSON path

A JSON path that points to an object that is used to define a table schema. The JSON path can be written in dot notation or bracket notation. The following operators are supported:

Description
Root element of a JSON object. This starts all path expressions
Wildcard character. Available anywhere a name or numeric are required in the JSON path.
Dot-notated child. Specifies a child field in a JSON object.

Description
Bracket-notated child. Specifies child field in a JSON object. Only a single child field can be specified.
Array index. Specifies the value of an array by index.

AWS Glue keeps track of the creation time, last update time, and version of your classifier.

Example Using a JSON classifier to pull records from an array

Suppose that your JSON data is an array of records. For example, the first few lines of your file might look like the following:

```
[
  {
    "type": "constituency",
    "id": "ocd-division\/country:us\/state:ak",
    "name": "Alaska"
  },
  {
    "type": "constituency",
    "id": "ocd-division\/country:us\/state:al\/cd:1",
    "name": "Alabama's 1st congressional district"
  },
  {
    "type": "constituency",
    "id": "ocd-division\/country:us\/state:al\/cd:2",
    "name": "Alabama's 2nd congressional district"
  },
  {
    "type": "constituency",
    "id": "ocd-division\/country:us\/state:al\/cd:3",
    "name": "Alabama's 3rd congressional district"
  },
  {
    "type": "constituency",
    "id": "ocd-division\/country:us\/state:al\/cd:4",
    "name": "Alabama's 4th congressional district"
  },
  {
    "type": "constituency",
    "id": "ocd-division\/country:us\/state:al\/cd:5",
    "name": "Alabama's 5th congressional district"
  },
  {
    "type": "constituency",
    "id": "ocd-division\/country:us\/state:al\/cd:6",
    "name": "Alabama's 6th congressional district"
  },
  {
    "type": "constituency",
    "id": "ocd-division\/country:us\/state:al\/cd:7",
    "name": "Alabama's 7th congressional district"
  },
  {
    "type": "constituency",
    "id": "ocd-division\/country:us\/state:ar\/cd:1",
    "name": "Arkansas's 1st congressional district"
  },
  {
    "type": "constituency",
    "id": "ocd-division\/country:us\/state:ar\/cd:2",
    "name": "Arkansas's 2nd congressional district"
  },
  {
    "type": "constituency",
    "id": "ocd-division\/country:us\/state:ar\/cd:3",
    "name": "Arkansas's 3rd congressional district"
  },
  {
    "type": "constituency",
    "id": "ocd-division\/country:us\/state:ar\/cd:4",
    "name": "Arkansas's 4th congressional district"
  }
]
```

When you run a crawler using the built-in JSON classifier, the entire file is used to define the schema. Because you don’t specify a JSON path, the crawler treats the data as one object, that is, just an array. For example, the schema might look like the following:

```

root
```

```
|-- record: array
```

However, to create a schema that is based on each record in the JSON array, create a custom JSON classifier and specify the JSON path as `$[*]`. When you specify this JSON path, the classifier interrogates all 12 records in the array to determine the schema. The resulting schema contains separate fields for each object, similar to the following example:

```
root
|-- type: string
|-- id: string
|-- name: string
```

Example Using a JSON classifier to examine only parts of a file

Suppose that your JSON data follows the pattern of the example JSON file `s3://awsglue-datasets/examples/us-legislators/all/areas.json` drawn from http://everypolitician.org/. Example objects in the JSON file look like the following:

```
{
  "type": "constituency",
  "id": "ocd-division\/country:us\/state:ak",
  "name": "Alaska"
}
{
  "type": "constituency",
  "identifiers": [
    {
      "scheme": "dmoz",
      "identifier": "Regional\/North_America\/United_States\/Alaska\/"
    },
    {
      "scheme": "freebase",
      "identifier": "\/m\/0hjy"
    },
    {
      "scheme": "fips",
      "identifier": "US02"
    },
    {
      "scheme": "quora",
      "identifier": "Alaska-state"
    },
    {
      "scheme": "britannica",
      "identifier": "place\/Alaska"
    },
    {
      "scheme": "wikidata",
      "identifier": "Q797"
    }
  ],
  "other_names": [
    {
      "lang": "en",
      "note": "multilingual",
      "name": "Alaska"
    },
    {
      "lang": "fr",
      "note": "multilingual",
      "name": "Alaska"
    },
    {
      "lang": "nov",
      "note": "multilingual",
      "name": "Alaska"
    }
  ],
  "id": "ocd-division\/country:us\/state:ak",
  "name": "Alaska"
}
```

When you run a crawler using the built-in JSON classifier, the entire file is used to create the schema. You might end up with a schema like this:

```
root
|-- type: string
|-- id: string
|-- name: string
|-- identifiers: array
|    |-- element: struct
|    |    |-- scheme: string
|    |    |-- identifier: string
|-- other_names: array
|    |-- element: struct
|    |    |-- lang: string
|    |    |-- note: string
|    |    |-- name: string
```

However, to create a schema using just the "`id`" object, create a custom JSON classifier and specify the JSON path as `$.id`. Then the schema is based on only the "`id`" field:

```

root
|-- record: string
```

The first few lines of data extracted with this schema look like this:

```

{"record": "ocd-division/country:us/state:ak"}
{"record": "ocd-division/country:us/state:al/cd:1"}
{"record": "ocd-division/country:us/state:al/cd:2"}
{"record": "ocd-division/country:us/state:al/cd:3"}
{"record": "ocd-division/country:us/state:al/cd:4"}
{"record": "ocd-division/country:us/state:al/cd:5"}
{"record": "ocd-division/country:us/state:al/cd:6"}
{"record": "ocd-division/country:us/state:al/cd:7"}
{"record": "ocd-division/country:us/state:ar/cd:1"}
{"record": "ocd-division/country:us/state:ar/cd:2"}
{"record": "ocd-division/country:us/state:ar/cd:3"}
{"record": "ocd-division/country:us/state:ar/cd:4"}
{"record": "ocd-division/country:us/state:as"}
{"record": "ocd-division/country:us/state:az/cd:1"}
{"record": "ocd-division/country:us/state:az/cd:2"}
{"record": "ocd-division/country:us/state:az/cd:3"}
{"record": "ocd-division/country:us/state:az/cd:4"}
{"record": "ocd-division/country:us/state:az/cd:5"}
{"record": "ocd-division/country:us/state:az/cd:6"}
{"record": "ocd-division/country:us/state:az/cd:7"}
```

To create a schema based on a deeply nested object, such as "`identifier`," in the JSON file, you can create a custom JSON classifier and specify the JSON path as

`$.identifiers[*].identifier`. Although the schema is similar to the previous example, it is based on a different object in the JSON file.

The schema looks like the following:

```
root
|-- record: string
```

Listing the first few lines of data from the table shows that the schema is based on the data in the "`identifier`" object:

```
{"record": "Regional/North_America/United_States/Alaska/"}
{"record": "/m/0hjy"}
{"record": "US02"}
{"record": "5879092"}
{"record": "4001016-8"}
{"record": "destination/alaska"}
{"record": "1116270"}
{"record": "139487266"}
{"record": "n79018447"}
{"record": "01490999-8dec-4129-8254-eef6e80fadc3"}
{"record": "Alaska-state"}
{"record": "place/Alaska"}
{"record": "Q797"}
{"record": "Regional/North_America/United_States/Alabama/"}
{"record": "/m/0gyh"}
{"record": "US01"}
{"record": "4829764"}
{"record": "4084839-5"}
{"record": "161950"}
{"record": "131885589"}
```

To create a table based on another deeply nested object, such as the "`name`" field in the "`other_names`" array in the JSON file, you can create a custom JSON classifier and specify the JSON path as `$.other_names[*].name`. Although the schema is similar to the previous example, it is based on a different object in the JSON file. The schema looks like the following:

```
root
```

```
|-- record: string
```

Listing the first few lines of data in the table shows that it is based on the data in the "`name`" object in the "`other_names`" array:

```

{"record": "Alaska"}
{"record": "Alaska"}
{"record": "Аляска"}
{"record": "Alaska"}
{"record": "Alaska"}
{"record": "Alaska"}
{"record": "Alaska"}
{"record": "Alaska"}
{"record": "Alaska"}
{"record": "######"}
{"record": "######"}
{"record": "######"}
{"record": "Alaska"}
{"record": "Alyaska"}
{"record": "Alaska"}
{"record": "Alaska"}
{"record": "Штат Аляска"}
{"record": "Аляска"}
{"record": "Alaska"}
{"record": "#######"}
```

Writing CSV custom classifiers

Custom CSV classifiers allows you to specify datatypes for each column in the custom csv classifier field. You can specify each column's datatype separated by a comma. By specifying datatypes, you can override the crawlers inferred datatypes and ensure data will be classified appropriately.

You can set the SerDe for processing CSV in the classifier, which will be applied in the Data Catalog.

When you create a custom classifier, you can also re-use the classifer for different crawlers.

- For csv files with only headers (no data), these files will be classified as UNKNOWN since not enough information is provided. If you specify that the CSV 'Has headings' in the *Column headings* option, and provide the datatypes, we can classify these files correctly.

You can use a custom CSV classifier to infer the schema of various types of CSV data. The custom attributes that you can provide for your classifier include delimiters, a CSV SerDe option, options about the header, and whether to perform certain validations on the data.

Custom classifier values in AWS Glue

When you define a CSV classifier, you provide the following values to AWS Glue to create the classifier. The classification field of this classifier is set to `csv`.

Classifier name

Name of the classifier.

CSV Serde

Sets the SerDe for processing CSV in the classifier, which will be applied in the Data Catalog. Options are Open CSV SerDe, Lazy Simple SerDe, and None. You can specify the None value when you want the crawler to do the detection.

Column delimiter

A custom symbol to denote what separates each column entry in the row. Provide a unicode character. If you cannot type your delimiter, you can copy and paste it. This works for printable characters, including those your system does not support (typically shown as □).

Quote symbol

A custom symbol to denote what combines content into a single column value. Must be different from the column delimiter. Provide a unicode character. If you cannot type your delimiter, you can copy and paste it. This works for printable characters, including those your system does not support (typically shown as □).

Column headings

Indicates the behavior for how column headings should be detected in the CSV file. If your custom CSV file has column headings, enter a comma-delimited list of the column headings.

Processing options: Allow files with single column

Enables the processing of files that contain only one column.

Processing options: Trim white space before identifying column values

Specifies whether to trim values before identifying the type of column values.

Custom datatypes - *optional*

Enter the custom datatype separated by a comma. Specifies the custom datatypes in the CSV file. The custom datatype must be a supported datatype. Supported datatypes are: “BINARY”, “BOOLEAN”, “DATE”, “DECIMAL”, “DOUBLE”, “FLOAT”, “INT”, “LONG”, “SHORT”, “STRING”, “TIMESTAMP”. Unsupported datatypes will display an error.

Creating classifiers using the AWS Glue console

A classifier determines the schema of your data. You can write a custom classifier and point to it from AWS Glue.

Creating classifiers

To add a classifier in the AWS Glue console, choose **Add classifier**. When you define a classifier, you supply values for the following:

- **Classifier name –** Provide a unique name for your classifier.
- **Classifier type –** The classification type of tables inferred by this classifier.
- **Last updated –** The last time this classifier was updated.

Classifier name

Provide a unique name for your classifier.

Classifier type

Choose the type of classifier to create.

Depending on the type of classifier you choose, configure the following properties for your classifier:

Grok

- **Classification**

  Describe the format or type of data that is classified or provide a custom label.
- **Grok pattern**

This is used to parse your data into a structured schema. The grok pattern is composed of named patterns that describe the format of your data store. You write this grok pattern using the named built-in patterns provided by AWS Glue and custom patterns you write and include in the **Custom patterns** field. Although grok debugger results might not match the results from AWS Glue exactly, we suggest that you try your pattern using some sample data with a grok debugger. You can find grok debuggers on the web. The named built-in patterns provided by AWS Glue are generally compatible with grok patterns that are available on the web.

Build your grok pattern by iteratively adding named patterns and check your results in a debugger. This activity gives you confidence that when the AWS Glue crawler runs your grok pattern, your data can be parsed.

- **Custom patterns**

  For grok classifiers, these are optional building blocks for the **Grok pattern** that you write. When built-in patterns cannot parse your data, you might need to write a custom pattern. These custom patterns are defined in this field and referenced in the **Grok pattern** field. Each custom pattern is defined on a separate line. Just like the built-in patterns, it consists of a named pattern definition that uses regular expression (regex) syntax.

  For example, the following has the name `MESSAGEPREFIX` followed by a regular expression definition to apply to your data to determine whether it follows the pattern.

  ```
  MESSAGEPREFIX .*-.*-.*-.*-.*
  ```

XML

- **Row tag**

  For XML classifiers, this is the name of the XML tag that defines a table row in the XML document. Type the name without angle brackets < >. The name must comply with XML rules for a tag.

  For more information, see Writing XML custom classifiers.

JSON

- **JSON path**

  For JSON classifiers, this is the JSON path to the object, array, or value that defines a row of the table being created. Type the name in either dot or bracket JSON syntax using AWS Glue supported operators.

  For more information, see the list of operators in Writing JSON custom classifiers.

CSV

- **Column delimiter**

  A single character or symbol to denote what separates each column entry in the row. Choose the delimiter from the list, or choose `Other` to enter a custom delimiter.
- **Quote symbol**

  A single character or symbol to denote what combines content into a single column value. Must be different from the column delimiter. Choose the quote symbol from the list, or choose `Other` to enter a custom quote character.
- **Column headings**

  Indicates the behavior for how column headings should be detected in the CSV file. You can choose `Has headings`, `No headings`, or `Detect headings`. If your custom CSV file has column headings, enter a comma-delimited list of the column headings.
- **Allow files with single column**

  To be classified as CSV, the data must have at least two columns and two rows of data. Use this option to allow the processing of files that contain only one column.
- **Trim whitespace before identifying column values**

  This option specifies whether to trim values before identifying the type of column values.
- **Custom datatype**

  (Optional) - Enter custom datatypes in a comma-delimited list. The supported datatypes are: “BINARY”, “BOOLEAN”, “DATE”, “DECIMAL”, “DOUBLE”, “FLOAT”, “INT”, “LONG”, “SHORT”, “STRING”, “TIMESTAMP”.

- **CSV Serde**

  (Optional) - A SerDe for processing CSV in the classifier, which will be applied in the Data Catalog. Choose from `Open CSV SerDe`, `Lazy Simple SerDe`, or `None`. You can specify the `None` value when you want the crawler to do the detection.

For more information, see Writing custom classifiers for diverse data formats.

Viewing classifiers

To see a list of all the classifiers that you have created, open the AWS Glue console at https://console.aws.amazon.com/glue/, and choose the **Classifiers** tab.

The list displays the following properties about each classifier:

- **Classifier** – The classifier name. When you create a classifier, you must provide a name for it.
- **Classification** – The classification type of tables inferred by this classifier.
- **Last updated** – The last time this classifier was updated.

Managing classifiers

From the **Classifiers** list in the AWS Glue console, you can add, edit, and delete classifiers. To see more details for a classifier, choose the classifier name in the list. Details include the information you defined when you created the classifier.

Configuring a crawler

A crawler accesses your data store, identifies metadata, and creates table definitions in the AWS Glue Data Catalog. The **Crawlers** pane in the AWS Glue console lists all the crawlers that you create. The list displays status and metrics from the last run of your crawler.

This topic contains the step-by-step process of configuring a crawler, covering essential aspects such as setting up the crawler's parameters, defining the data sources to crawl, setting up security, and managing the crawled data.

Topics

- Step 1: Set crawler properties
- Step 2: Choose data sources and classifiers

- Step 3: Configure security settings
- Step 4: Set output and scheduling
- Step 5: Review and create

Step 1: Set crawler properties

To configure a crawler

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/. Choose **Crawlers** in the navigation pane.
2. Choose **Create crawler**, and follow the instructions in the **Add crawler** wizard. The wizard will guide you the steps required to create a crawler. If you want to add custom calssifiers to define the schema, see Defining and managing classifiers.
3. Enter a name for your crawler and description (optional). Optionally, you can tag your crawler with a **Tag key** and optional **Tag value**. Once created, tag keys are read-only. Use tags on some resources to help you organize and identify them. For more information, see AWS tags in AWS Glue.

   Name

      Name may contain letters (A-Z), numbers (0-9), hyphens (-), or underscores (_), and can be up to 255 characters long.

   Description

      Descriptions can be up to 2048 characters long.

   Tags

      Use tags to organize and identify your resources. For more information, see the following:

      - AWS tags in AWS Glue

Step 2: Choose data sources and classifiers

Next, configure the data sources and classifiers for the crawler.

For more information about supported data sources, see Supported data sources for crawling.

Data source configuration

Select the appropriate option for **Is your data already mapped to AWS Glue tables?** choose 'Not yet' or 'Yes'. By default, 'Not yet' is selected.

The crawler can access data stores directly as the source of the crawl, or it can use existing tables in the Data Catalog as the source. If the crawler uses existing catalog tables, it crawls the data stores that are specified by those catalog tables.

- Not yet: Select one or more data sources to be crawled. A crawler can crawl multiple data stores of different types (Amazon S3, JDBC, and so on).

  You can configure only one data store at a time. After you have provided the connection information and include paths and exclude patterns, you then have the option of adding another data store.
- Yes: Select existing tables from your AWS Glue Data Catalog. The catalog tables specify the data stores to crawl. The crawler can crawl only catalog tables in a single run; it can't mix in other source types.

  A common reason to specify a catalog table as the source is when you create the table manually (because you already know the structure of the data store) and you want a crawler to keep the table updated, including adding new partitions. For a discussion of other reasons, see Updating manually created Data Catalog tables using crawlers.

  When you specify existing tables as the crawler source type, the following conditions apply:
  - Database name is optional.
  - Only catalog tables that specify Amazon S3, Amazon DynamoDB, or Delta Lake data stores are permitted.
  - No new catalog tables are created when the crawler runs. Existing tables are updated as needed, including adding new partitions.
  - Deleted objects found in the data stores are ignored; no catalog tables are deleted. Instead, the crawler writes a log message. (`SchemaChangePolicy.DeleteBehavior=LOG`)
  - The crawler configuration option to create a single schema for each Amazon S3 path is enabled by default and cannot be disabled. (`TableGroupingPolicy=CombineCompatibleSchemas`) For more information, see Creating a single schema for each Amazon S3 include path.
  - You can't mix catalog tables as a source with any other source types (for example Amazon S3 or Amazon DynamoDB).

To use Delta tables, first create a Delta table using Athena DDL or the AWS Glue API.

Using Athena, set the location to your Amazon S3 folder and the table type to 'DELTA'.

```
CREATE EXTERNAL TABLE database_name.table_name
LOCATION 's3://bucket/folder/'
TBLPROPERTIES ('table_type' = 'DELTA')
```

Using the AWS Glue API, specify the table type within the table parameters map. The table parameters need to include the following key/value pair. For more information on how to create a table, see Boto3 documentation for create_table .

```
{
    "table_type":"delta"
}
```

Data sources

Select or add the list of data sources to be scanned by the crawler.

(Optional) If you choose JDBC as the data source, you can use your own JDBC drivers when specifying the Connection access where the driver info is stored.

Include path

When evaluating what to include or exclude in a crawl, a crawler starts by evaluating the required include path. For Amazon S3, MongoDB, MongoDB Atlas, Amazon DocumentDB (with MongoDB compatibility), and relational data stores, you must specify an include path.

For an Amazon S3 data store

Choose whether to specify a path in this account or in a different account, and then browse to choose an Amazon S3 path.

For Amazon S3 data stores, include path syntax is `bucket-name/folder-name/file-name.ext`. To crawl all objects in a bucket, you specify just the bucket name in the include path. The exclude pattern is relative to the include path

For a Delta Lake data store

Specify one or more Amazon S3 paths to Delta tables as s3://*bucket*/*prefix*/*object*.

For an Iceberg or Hudi data store

Specify one or more Amazon S3 paths that contain folders with Iceberg or Hudi table metadata as s3://*bucket*/*prefix*.

For Iceberg and Hudi data stores, the Iceberg/Hudi folder may be located in a child folder of the root folder. The crawler will scan all folders underneath a path for a Hudi folder.

For a JDBC data store

Enter *<database>*/*<schema>*/*<table>* or *<database>*/*<table>*, depending on the database product. Oracle Database and MySQL don't support schema in the path. You can substitute the percent (%) character for *<schema>* or *<table>*. For example, for an Oracle database with a system identifier (SID) of `orcl`, enter `orcl/%` to import all tables to which the user named in the connection has access.

> ⚠ **Important**
>
> This field is case-sensitive.

> ⓘ **Note**
>
> If you choose to bring in your own JDBC driver versions, AWS Glue crawlers will consume resources in AWS Glue jobs and Amazon S3 buckets to ensure your provided driver are run in your environment. The additional usage of resources will be reflected in your account. Drivers are limited to the properties described in Adding an AWS Glue connection.

For a MongoDB, MongoDB Atlas, or Amazon DocumentDB data store

For MongoDB, MongoDB Atlas, and Amazon DocumentDB (with MongoDB compatibility), the syntax is `database/collection`.

For JDBC data stores, the syntax is either `database-name/schema-name/table-name` or `database-name/table-name`. The syntax depends on whether the database engine supports schemas within a database. For example, for database engines such as MySQL or Oracle, don't specify a `schema-name` in your include path. You can substitute the percent sign (`%`) for a schema or table in the include path to represent all schemas or all tables in a database. You cannot substitute the percent sign (%) for database in the include path.

Maximum transversal depth (for Iceberg or Hudi data stores only)

Defines the maximum depth of the Amazon S3 path that the crawler can traverse to discover the Iceberg or Hudi metadata folder in your Amazon S3 path. The purpose of this parameter is to limit the crawler run time. The default value is 10 and the maximum is 20.

Exclude patterns

These enable you to exclude certain files or tables from the crawl. The exclude path is relative to the include path. For example, to exclude a table in your JDBC data store, type the table name in the exclude path.

A crawler connects to a JDBC data store using an AWS Glue connection that contains a JDBC URI connection string. The crawler only has access to objects in the database engine using the JDBC user name and password in the AWS Glue connection. *The crawler can only create tables that it can access through the JDBC connection.* After the crawler accesses the database engine with the JDBC URI, the include path is used to determine which tables in the database engine are created in the Data Catalog. For example, with MySQL, if you specify an include path of `MyDatabase/%`, then all tables within `MyDatabase` are created in the Data Catalog. When accessing Amazon Redshift, if you specify an include path of `MyDatabase/%`, then all tables within all schemas for database `MyDatabase` are created in the Data Catalog. If you specify an include path of `MyDatabase/MySchema/%`, then all tables in database `MyDatabase` and schema `MySchema` are created.

After you specify an include path, you can then exclude objects from the crawl that your include path would otherwise include by specifying one or more Unix-style `glob` exclude patterns. These patterns are applied to your include path to determine which objects are excluded. These patterns are also stored as a property of tables created by the crawler. AWS Glue PySpark extensions, such as `create_dynamic_frame.from_catalog`, read the table properties and exclude objects defined by the exclude pattern.

AWS Glue supports the following `glob` patterns in the exclude pattern.

Exclude pattern	Description
`*.csv`	Matches an Amazon S3 path that represents an object name in the current folder ending in `.csv`
`*.*`	Matches all object names that contain a dot
`*.{csv,avro}`	Matches object names ending with `.csv` or `.avro`
`foo.?`	Matches object names starting with `foo.` that are followed by a single character extension
`myfolder/*`	Matches objects in one level of subfolder from `myfolder`, such as `/myfolder/mysource`
`myfolder/*/*`	Matches objects in two levels of subfolders from `myfolder`, such as `/myfolder/mysource/data`
`myfolder/**`	Matches objects in all subfolders of `myfolder`, such as `/myfolder/mysource/mydata` and `/myfolder/mysource/data`
`myfolder**`	Matches subfolder `myfolder` as well as files below `myfolder`, such as `/myfolder` and `/myfolder/mydata.txt`
`Market*`	Matches tables in a JDBC database with names that begin with `Market`, such as `Market_us` and `Market_fr`

AWS Glue interprets `glob` exclude patterns as follows:

- The slash (/) character is the delimiter to separate Amazon S3 keys into a folder hierarchy.
- The asterisk (*) character matches zero or more characters of a name component without crossing folder boundaries.
- A double asterisk (`**`) matches zero or more characters crossing folder or schema boundaries.
- The question mark (?) character matches exactly one character of a name component.
- The backslash (\) character is used to escape characters that otherwise can be interpreted as special characters. The expression `\\` matches a single backslash, and `\{` matches a left brace.
- Brackets `[ ]` create a bracket expression that matches a single character of a name component out of a set of characters. For example, `[abc]` matches `a`, `b`, or `c`. The hyphen (`-`) can be used to specify a range, so `[a-z]` specifies a range that matches from `a` through `z` (inclusive). These forms can be mixed, so [`abce-g`] matches `a`, `b`, `c`, `e`, `f`, or `g`. If the character after the bracket (`[`) is an exclamation point (`!`), the bracket expression is negated. For example, `[!a-c]` matches any character except `a`, `b`, or `c`.

  Within a bracket expression, the `*`, `?`, and `\` characters match themselves. The hyphen (`-`) character matches itself if it is the first character within the brackets, or if it's the first character after the `!` when you are negating.
- Braces (`{ }`) enclose a group of subpatterns, where the group matches if any subpattern in the group matches. A comma (`,`) character is used to separate the subpatterns. Groups cannot be nested.
- Leading period or dot characters in file names are treated as normal characters in match operations. For example, the `*` exclude pattern matches the file name `.hidden`.

Example Amazon S3 exclude patterns

Each exclude pattern is evaluated against the include path. For example, suppose that you have the following Amazon S3 directory structure:

```
/mybucket/myfolder/
   departments/
      finance.json
      market-us.json
      market-emea.json
      market-ap.json
   employees/
      hr.json
```

```
        john.csv
        jane.csv
        juan.txt
```

Given the include path `s3://mybucket/myfolder/`, the following are some sample results for exclude patterns:

Exclude pattern	Results
`departments/**`	Excludes all files and folders below `departments` and includes the `employees` folder and its files
`departments/market*`	Excludes `market-us.json`, `market-emea.json`, and `market-ap.json`
`**.csv`	Excludes all objects below `myfolder` that have a name ending with `.csv`
`employees/*.csv`	Excludes all `.csv` files in the `employees` folder

Example Excluding a subset of Amazon S3 partitions

Suppose that your data is partitioned by day, so that each day in a year is in a separate Amazon S3 partition. For January 2015, there are 31 partitions. Now, to crawl data for only the first week of January, you must exclude all partitions except days 1 through 7:

```
2015/01/{[!0],0[8-9]}**, 2015/0[2-9]/**, 2015/1[0-2]/**
```

Take a look at the parts of this glob pattern. The first part, `2015/01/{[!0],0[8-9]}**`, excludes all days that don't begin with a "0" in addition to day 08 and day 09 from month 01 in year 2015. Notice that "**" is used as the suffix to the day number pattern and crosses folder boundaries to lower-level folders. If "*" is used, lower folder levels are not excluded.

The second part, `2015/0[2-9]/**`, excludes days in months 02 to 09, in year 2015.

The third part, `2015/1[0-2]/**`, excludes days in months 10, 11, and 12, in year 2015.

Example JDBC exclude patterns

Suppose that you are crawling a JDBC database with the following schema structure:

```
MyDatabase/MySchema/
   HR_us
   HR_fr
   Employees_Table
   Finance
   Market_US_Table
   Market_EMEA_Table
   Market_AP_Table
```

Given the include path `MyDatabase/MySchema/%`, the following are some sample results for exclude patterns:

Exclude pattern	Results
HR*	Excludes the tables with names that begin with HR
`Market_*`	Excludes the tables with names that begin with `Market_`
`**_Table`	Excludes all tables with names that end with `_Table`

Additional crawler source parameters

Each source type requires a different set of additional parameters.

Connection

Select or add an AWS Glue connection. For information about connections, see Connecting to data.

Additional metadata - optional (for JDBC data stores)

Select additional metadata properties for the crawler to crawl.

- Comments: Crawl associated table level and column level comments.
- Raw types: Persist the raw datatypes of the table columns in additional metadata. As a default behavior, the crawler translates the raw datatypes to Hive-compatible types.

JDBC Driver Class name - optional (for JDBC data stores)

Type a custom JDBC driver class name for the crawler to connect to the data source:

- Postgres: org.postgresql.Driver
- MySQL: com.mysql.jdbc.Driver, com.mysql.cj.jdbc.Driver
- Redshift: com.amazon.redshift.jdbc.Driver, com.amazon.redshift.jdbc42.Driver
- Oracle: oracle.jdbc.driver.OracleDriver
- SQL Server: com.microsoft.sqlserver.jdbc.SQLServerDriver

JDBC Driver S3 Path - optional (for JDBC data stores)

Choose an existing Amazon S3 path to a `.jar` file. This is where the `.jar` file will be stored when using a custom JDBC driver for the crawler to connect to the data source.

Enable data sampling (for Amazon DynamoDB, MongoDB, MongoDB Atlas, and Amazon DocumentDB data stores only)

Select whether to crawl a data sample only. If not selected the entire table is crawled. Scanning all the records can take a long time when the table is not a high throughput table.

Create tables for querying (for Delta Lake data stores only)

Select how you want to create the Delta Lake tables:

- Create Native tables: Allow integration with query engines that support querying of the Delta transaction log directly.
- Create Symlink tables: Create a symlink manifest folder with manifest files partitioned by the partition keys, based on the specified configuration parameters.

Scanning rate - optional (for DynamoDB data stores only)

Specify the percentage of the DynamoDB table Read Capacity Units to use by the crawler. Read capacity units is a term defined by DynamoDB, and is a numeric value that acts as rate limiter for the number of reads that can be performed on that table per second. Enter a value between 0.1 and 1.5. If not specified, defaults to 0.5 for provisioned tables and 1/4 of maximum configured capacity for on-demand tables. Note that only provisioned capacity mode should be used with AWS Glue crawlers.

> **Note**
>
> For DynamoDB data stores, set the provisioned capacity mode for processing reads and writes on your tables. The AWS Glue crawler should not be used with the on-demand capacity mode.

Network connection - optional (for Amazon S3, Delta, Iceberg, Hudi and Catalog target data stores)

Optionally include a Network connection to use with this Amazon S3 target. Note that each crawler is limited to one Network connection so any other Amazon S3 targets will also use the same connection (or none, if left blank).

For information about connections, see Connecting to data.

Sample only a subset of files and Sample size (for Amazon S3 data stores only)

Specify the number of files in each leaf folder to be crawled when crawling sample files in a dataset. When this feature is turned on, instead of crawling all the files in this dataset, the crawler randomly selects some files in each leaf folder to crawl.

The sampling crawler is best suited for customers who have previous knowledge about their data formats and know that schemas in their folders do not change. Turning on this feature will significantly reduce crawler runtime.

A valid value is an integer between 1 and 249. If not specified, all the files are crawled.

Subsequent crawler runs

This field is a global field that affects all Amazon S3 data sources.

- Crawl all sub-folders: Crawl all folders again with every subsequent crawl.
- Crawl new sub-folders only: Only Amazon S3 folders that were added since the last crawl will be crawled. If the schemas are compatible, new partitions will be added to existing tables. For more information, see the section called “Scheduling incremental crawls”.
- Crawl based on events: Rely on Amazon S3 events to control what folders to crawl. For more information, see Accelerating crawls using Amazon S3 event notifications.

Custom classifiers - optional

Define custom classifiers before defining crawlers. A classifier checks whether a given file is in a format the crawler can handle. If it is, the classifier creates a schema in the form of a `StructType` object that matches that data format.

For more information, see Defining and managing classifiers.

Step 3: Configure security settings

IAM role

The crawler assumes this role. It must have permissions similar to the AWS managed policy `AWSGlueServiceRole`. For Amazon S3 and DynamoDB sources, it must also have permissions to access the data store. If the crawler reads Amazon S3 data encrypted with AWS Key Management Service (AWS KMS), then the role must have decrypt permissions on the AWS KMS key.

For an Amazon S3 data store, additional permissions attached to the role would be similar to the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "s3:GetObject",
        "s3:PutObject"
      ],
      "Resource": [
        "arn:aws:s3:::bucket/object*"
      ]
    }
  ]
}
```

For an Amazon DynamoDB data store, additional permissions attached to the role would be similar to the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "dynamodb:DescribeTable",
        "dynamodb:Scan"
      ],
      "Resource": [
        "arn:aws:dynamodb:*:111122223333:table/table-name*"
      ]
    }
  ]
}
```

In order to add your own JDBC driver, additional permissions need to be added.

- Grant permissions for the following job actions: `CreateJob`, `DeleteJob`, `GetJob`, `GetJobRun`, `StartJobRun`.
- Grant permissions for Amazon S3 actions: `s3:DeleteObjects`, `s3:GetObject`, `s3:ListBucket`, `s3:PutObject`.

> **Note**
>
> The `s3:ListBucket` is not needed if the Amazon S3 bucket policy is disabled.

- Grant service principal access to bucket/folder in the Amazon S3 policy.

Example Amazon S3 policy:

JSON

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "VisualEditor0",
            "Effect": "Allow",
            "Action": [
                "s3:PutObject",
                "s3:GetObject",
                "s3:ListBucket",
                "s3:DeleteObject"
            ],
            "Resource": [
                "arn:aws:s3:::amzn-s3-demo-bucket/driver-parent-folder/
driver.jar",
                "arn:aws:s3:::amzn-s3-demo-bucket"
            ]
        }
    ]
}
```

AWS Glue creates the following folders (`_crawler` and `_glue_job_crawler` at the same level as the JDBC driver in your Amazon S3 bucket. For example, if the driver path is `<s3-path/driver_folder/driver.jar>`, then the following folders will be created if they do not already exist:

- <s3-path/driver_folder/_crawler>
- <s3-path/driver_folder/_glue_job_crawler>

Optionally, you can add a security configuration to a crawler to specify at-rest encryption options.

For more information, see Step 2: Create an IAM role for AWS Glue and Identity and access management for AWS Glue.

Lake Formation configuration - optional

Allow the crawler to use Lake Formation credentials for crawling the data source.

Checking **Use Lake Formation credentials for crawling S3 data source** will allow the crawler to use Lake Formation credentials for crawling the data source. If the data source belongs to another account, you must provide the registered account ID. Otherwise, the crawler will crawl only those data sources associated to the account. Only applicable to Amazon S3 and Data Catalog data sources.

Security configuration - optional

Settings include security configurations. For more information, see the following:

- Encrypting data written by AWS Glue

> **Note**
>
> Once a security configuration has been set on a crawler, you can change, but you cannot remove it. To lower the level of security on a crawler, explicitly set the security feature to `DISABLED` within your configuration, or create a new crawler.

Step 4: Set output and scheduling

Output configuration

Options include how the crawler should handle detected schema changes, deleted objects in the data store, and more. For more information, see Customizing crawler behavior

Crawler schedule

You can run a crawler on demand or define a time-based schedule for your crawlers and jobs in AWS Glue. The definition of these schedules uses the Unix-like cron syntax. For more information, see Scheduling a crawler.

Step 5: Review and create

Review the crawler settings you configured, and create the crawler.

Scheduling a crawler

You can run an AWS Glue crawler on demand or on a regular schedule. When you set up a crawler based on a schedule, you can specify certain constraints, such as the frequency of the crawler runs, which days of the week it runs, and at what time. You can create these custom schedules in *cron* format. For more information, see cron in Wikipedia.

When setting up a crawler schedule, you should consider the features and limitations of cron. For example, if you choose to run your crawler on day 31 each month, keep in mind that some months don't have 31 days.

Topics

- Create a crawler schedule
- Create a schedule for an existing crawler

Create a crawler schedule

You can create a schedule for the crawler using the AWS Glue console or AWS CLI.

AWS Management Console

1. Sign in to the AWS Management Console, and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. Choose **Crawlers** in the navigation pane.
3. Follow steps 1-3 in the Configuring a crawler section.
4. In Step 4: Set output and scheduling, choose a **Crawler schedule** to set the frequency of the run. You can choose the crawler to run hourly, daily, weekly, monthly or define custom schedule using cron expressions.

   A cron expression is a string representing a schedule pattern, consisting of 6 fields separated by spaces: * * * * * <minute> <hour> <day of month> <month> <day of week> <year>

   For example, to run a task every day at midnight, the cron expression is: 0 0 * * ? *

   For more information, see Cron expressions.
5. Review the crawler settings you configured, and create the crawler to run on a schedule.

AWS CLI

```
aws glue create-crawler
 --name myCrawler \
 --role AWSGlueServiceRole-myCrawler  \
 --targets '{"S3Targets":[{Path="s3://amzn-s3-demo-bucket/"}]}' \
 --schedule cron(15 12 * * ? *)
```

For more information about using cron to schedule jobs and crawlers, see Time-based schedules for jobs and crawlers.

Create a schedule for an existing crawler

Follow these steps to set up a recurring schedule for an existing crawler.

AWS Management Console

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. Choose **Crawlers** in the navigation pane.
3. Choose a crawler that you want to schedule from the available list.
4. Choose **Edit** from the **Actions menu.**
5. Scroll down to **Step 4: Set output and scheduling**, and choose **Edit**.
6. Update your crawler schedule under **Crawler schedule**.
7. Choose **Update**.

AWS CLI

Use the following CLI command to update an existing crawler configuration:

```
aws glue update-crawler-schedule
   --crawler-name myCrawler
   --schedule cron(15 12 * * ? *)
```

Viewing crawler results and details

After the crawler runs successfully, it creates table definitions in the Data Catalog. Choose **Tables** in the navigation pane to see the tables that were created by your crawler in the database that you specified.

You can view information related to the crawler itself as follows:

- The **Crawlers** page on the AWS Glue console displays the following properties for a crawler:

Property	Description
Name	When you create a crawler, you must give it a unique name.
Status	A crawler can be ready, starting, stopping, scheduled, or schedule paused. A running

Property	Description
	crawler progresses from starting to stopping. You can resume or pause a schedule attached to a crawler.
Schedule	You can choose to run your crawler on demand or choose a frequency with a schedule. For more information about scheduling a crawler, see Scheduling a crawler.
Last run	The date and time of the last time the crawler was run.
Log	Links to any available logs from the last run of the crawler.
Tables changes from last run	The number of tables in the AWS Glue Data Catalog that were updated by the latest run of the crawler.

- To view the history of a crawler, choose **Crawlers** in the navigation pane to see the crawlers you created. Choose a crawler from the list of available crawlers. You can view the crawler properties and view the crawler history in the **Crawler runs** tab.

  The Crawler runs tab displays information about each time the crawler ran, including **Start time (UTC), End time (UTC)**, **Duration**, **Status**, **DPU hours**, and **Table changes**.

  The Crawler runs tab displays only the crawls that have occurred since the launch date of the crawler history feature, and only retains up to 12 months of crawls. Older crawls will not be returned.

- To see additional information, choose a tab in the crawler details page. Each tab will display information related to the crawler.
  - **Schedule**: Any schedules created for the crawler will be visible here.
  - **Data sources**: All data sources scanned by the crawler will be visible here.
  - **Classifiers**: All classifiers assigned to the crawler will be visible here.
  - **Tags**: Any tags created and assigned to an AWS resource will be visible here.

Parameters set on Data Catalog tables by crawler

These table properties are set by AWS Glue crawlers. We expect users to consume the `classification` and `compressionType` properties. Other properties, including table size estimates, are used for internal calculations, and we do not guarantee their accuracy or applicability to customer use cases. Changing these parameters may alter the behavior of the crawler, we do not support this workflow.

Property key	Property value
`UPDATED_BY_CRAWLER`	Name of crawler performing update.
`connectionName`	The name of the connection in the Data Catalog for the crawler used to connect the to the data store.
`recordCount`	Estimate count of records in table, based on file sizes and headers.
`skip.header.line.count`	Rows skipped to skip header. Set on tables classified as CSV.
`CrawlerSchemaSerializerVersion`	For internal use
`classification`	Format of data, inferred by crawler. For more information about data formats supported by AWS Glue crawlers see the section called “Built-in classifiers”.
`CrawlerSchemaDeserializerVersion`	For internal use
`sizeKey`	Combined size of files in table crawled.
`averageRecordSize`	Average size of row in table, in bytes.

Property key	Property value
`compressionType`	Type of compression used on data in the table. For more information about compression types supported by AWS Glue crawlers see the section called "Built-in classifiers".
`typeOfData`	`file`, `table` or `view`.
`objectCount`	Number of objects under Amazon S3 path for table.

These additional table properties are set by AWS Glue crawlers for Snowflake data stores.

Property key	Property value
`aws:RawTableLastAltered`	Records the last altered timestamp of the Snowflake table.
`ViewOriginalText`	View SQL statement.
`ViewExpandedText`	View SQL statement encoded in Base64 format.
`ExternalTable:S3Location`	Amazon S3 location of the Snowflake external table.
`ExternalTable:FileFormat`	Amazon S3 file format of the Snowflake external table.

These additional table properties are set by AWS Glue crawlers for JDBC-type data stores such as Amazon Redshift, Microsoft SQL Server, MySQL, PostgreSQL, and Oracle.

Property key	Property value
`aws:RawType`	When a crawler store the data in the Data Catalog it translates the datatypes to Hive-compatible types, which many times causes the information on the native datatype to be lost. The crawler outputs the `aws:RawType` parameter to provide the native-level datatype.
`aws:RawColumnComment`	If a comment is associated with a column in the database, the crawler outputs the corresponding comment in the catalog table. The comment string is truncated to 255 bytes. Comments are not supported for Microsoft SQL Server.
`aws:RawTableComment`	If a comment is associated with a table in the database, the crawler outputs corresponding comment in the catalog table. The comment string is truncated to 255 bytes. Comments are not supported for Microsoft SQL Server.

Customizing crawler behavior

When you configure an AWS Glue crawler, you have several options for defining the behavior of your crawler.

- **Incremental crawls –** You can configure a crawler to run incremental crawls to add only new partitions to the table schema.
- **Partition indexes –** A crawler creates partition indexes for Amazon S3 and Delta Lake targets by default to provide efficient lookup for specific partitions.
- **Accelerate crawl time by using Amazon S3 events –** You can configure a crawler to use Amazon S3 events to identify the changes between two crawls by listing all the files from the subfolder which triggered the event instead of listing the full Amazon S3 or Data Catalog target.
- **Handling schema changes –** You can prevent a crawlers from making any schema changes to the existing schema. You can use the AWS Management Console or the AWS Glue API to configure how your crawler processes certain types of changes.
- **A single schema for multiple Amazon S3 paths –** You can configure a crawler to create a single schema for each S3 path if the data is compatible.

- **Table location and partitioning levels** – The table level crawler option provides you the flexibility to tell the crawler where the tables are located, and how you want partitions created.
- **Table threshold** – You can specify the maximum number of tables the crawler is allowed to create by specifying a table threshold.
- **AWS Lake Formation credentials** – You can configure a crawler to use Lake Formation credentials to access an Amazon S3 data store or a Data Catalog table with an underlying Amazon S3 location within the same AWS account or another AWS account.

For more information about using the AWS Glue console to add a crawler, see Configuring a crawler.

Topics

- Scheduling incremental crawls for adding new partitions
- Generating partition indexes
- Preventing a crawler from changing an existing schema
- Creating a single schema for each Amazon S3 include path
- Specifying the table location and partitioning level
- Specifying the maximum number of tables the crawler is allowed to create
- Configuring a crawler to use Lake Formation credentials
- Accelerating crawls using Amazon S3 event notifications

Scheduling incremental crawls for adding new partitions

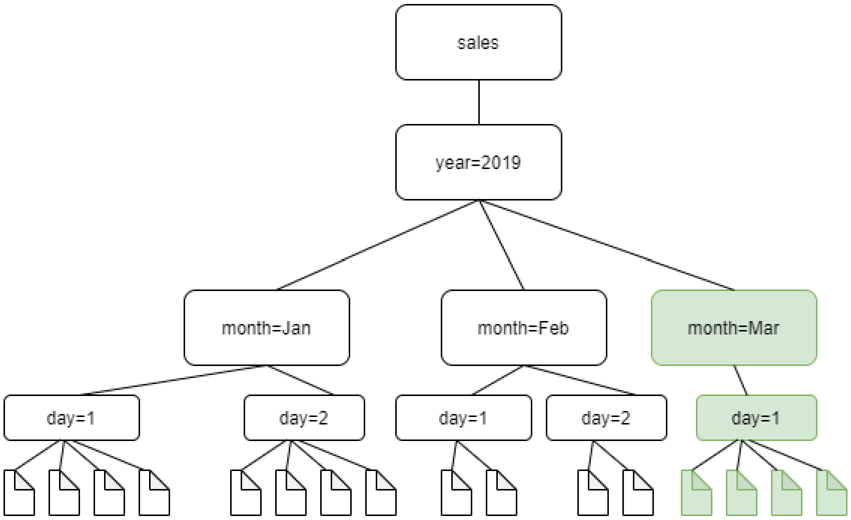
You can configure an AWS Glue crawler run incremental crawls to add only new partitions to the table schema. When the crawler runs for the first time, it performs a full crawl to processes the entire data source to record the complete schema and all existing partitions in the AWS Glue Data Catalog.

Subsequent crawls after the initial full crawl will be incremental, where the crawler identifies and adds only the new partitions that have been introduced since the previous crawl. This approach results in faster crawl times, as the crawler no longer needs to process the entire data source for each run, but instead focuses only on the new partitions.

Note

Incremental crawls don't detect modifications or deletions of existing partitions. This configuration is best suited for data sources with a stable schema. If a one-time major schema change occurs, it is advisable to temporarily set the crawler to perform a full crawl to capture the new schema accurately, and then switch back to incremental crawling mode.

The following diagram shows that with the incremental crawl setting enabled, the crawler will only detect and add the newly added folder, month=March, to the catalog.



Follow these steps to update your crawler to perform incremental crawls:

AWS Management Console

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. Choose **Crawlers** under the **Data Catalog**.
3. Choose a crawler that you want to set up to crawl incrementally.
4. Choose **Edit**.
5. Choose **Step 2. Choose data sources and classifiers**.
6. Choose the data source that you want to incrementally crawl.
7. Choose **Edit**.
8. Choose **Crawl new sub-folders only** under **Subsequent crawler runs**.
9. Choose **Update**.

To create a schedule for a crawler, see the section called "Scheduling a crawler".

AWS CLI

```
aws glue update-crawler \
 --name myCrawler \
 --recrawl-policy RecrawlBehavior=CRAWL_NEW_FOLDERS_ONLY \
 --schema-change-policy UpdateBehavior=LOG,DeleteBehavior=LOG
```

Notes and restrictions

When this option is turned on, you can't change the Amazon S3 target data stores when editing the crawler. This option affects certain crawler configuration settings. When turned on, it forces the update behavior and delete behavior of the crawler to `LOG`. This means that:

- If it discovers objects where schemas are not compatible, the crawler will not add the objects in the Data Catalog, and adds this detail as a log in CloudWatch Logs.
- It will not update deleted objects in the Data Catalog.

Generating partition indexes

The Data Catalog supports creating partition indexes to provide efficient lookup for specific partitions. For more information, see Creating partition indexes. The AWS Glue crawler creates partition indexes for Amazon S3 and Delta Lake targets by default.

AWS Management Console

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. Choose **Crawlers** under the **Data Catalog**.
3. When you define a crawler, the option to **Create partition indexes automatically** is enabled by default under **Advanced options** on the **Set output and scheduling** page.

   To disable this option, you can unselect the checkbox **Create partition indexes automatically** in the console.
4. Complete the crawler configuration and choose **Create crawler**.

AWS CLI

You can also disable this option by using the AWS CLI, set the `CreatePartitionIndex` in the `configuration` parameter. The default value is true.

```
aws glue update-crawler \
    --name myCrawler \
    --configuration '{"Version": 1.0, "CreatePartitionIndex": false }'
```

Usage notes for partition indexes

- Tables created by the crawler do not have the variable `partition_filtering.enabled` by default. For more information, see AWS Glue partition indexing and filtering.
- Creating partition indexes for encrypted partitions is not supported.

Preventing a crawler from changing an existing schema

You can prevent AWS Glue crawlers from making any schema changes to the Data Catalog when they run. By default, crawlers updates the schema in the Data Catalog to match the data source being crawled. However, in some cases, you may want to prevent the Crawler from modifying the existing schema, especially if you have transformed or cleaned the data and don't want the original schema to overwrite the changes.

Follow these steps to configure your crawler not to overwrite the existing schema in a table definition.

AWS Management Console

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. Choose **Crawlers** under the **Data Catalog**.
3. Choose a crawler from the list, and choose **Edit**.
4. Choose **step 4, Set output and scheduling**.
5. Under **Advance options**, choose **Add new columns only** or **Ignore the change and don't update the table in the Data Catalog**.
6. You can also set a configuration option to **Update all new and existing partitions with metadata from the table**. This sets partition schemas to inherit from the table.

7. Choose **Update**.

AWS CLI

The following example shows how to configure a crawler to not change existing schema, only add new columns:

```
aws glue update-crawler \
  --name myCrawler \
  --configuration '{"Version": 1.0, "CrawlerOutput": {"Tables":
 {"AddOrUpdateBehavior": "MergeNewColumns"}}}'
```

The following example shows how to configure a crawler to not change the existing schema, and not add new columns:

```
aws glue update-crawler \
  --name myCrawler \
  --schema-change-policy UpdateBehavior=LOG \
  --configuration '{"Version": 1.0, "CrawlerOutput": {"Partitions":
 { "AddOrUpdateBehavior": "InheritFromTable" }}}'
```

API

If you don't want a table schema to change at all when a crawler runs, set the schema change policy to `LOG`.

When you configure the crawler using the API, set the following parameters:

- Set the `UpdateBehavior` field in `SchemaChangePolicy` structure to `LOG`.
- Set the `Configuration` field with a string representation of the following JSON object in the crawler API; for example:

  ```
  {
     "Version": 1.0,
     "CrawlerOutput": {
        "Partitions": { "AddOrUpdateBehavior": "InheritFromTable" }
     }
  }
  ```

Creating a single schema for each Amazon S3 include path

By default, when a crawler defines tables for data stored in Amazon S3, it considers both data compatibility and schema similarity. Data compatibility factors that it considers include whether the data is of the same format (for example, JSON), the same compression type (for example, GZIP), the structure of the Amazon S3 path, and other data attributes. Schema similarity is a measure of how closely the schemas of separate Amazon S3 objects are similar.

To help illustrate this option, suppose that you define a crawler with an include path `s3://amzn-s3-demo-bucket/table1/`. When the crawler runs, it finds two JSON files with the following characteristics:

- **File 1** – `S3://amzn-s3-demo-bucket/table1/year=2017/data1.json`
- *File content* – `{“A”: 1, “B”: 2}`
- *Schema* – `A:int, B:int`

- **File 2** – `S3://amzn-s3-demo-bucket/table1/year=2018/data2.json`
- *File content* – `{“C”: 3, “D”: 4}`
- *Schema* – `C: int, D: int`

By default, the crawler creates two tables, named `year_2017` and `year_2018` because the schemas are not sufficiently similar. However, if the option **Create a single schema for each S3 path** is selected, and if the data is compatible, the crawler creates one table. The table has the schema `A:int,B:int,C:int,D:int` and `partitionKey` `year:string`.

AWS Management Console

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. Choose **Crawlers** under the **Data Catalog**.
3. When you configure a new crawler, under **Output and scheduling** , select the option **Create a single schema for each S3 path** under Advance options.

AWS CLI

You can configure a crawler to `CombineCompatibleSchemas` into a common table definition when possible. With this option, the crawler still considers data compatibility, but ignores the similarity of the specific schemas when evaluating Amazon S3 objects in the specified include path.

When you configure the crawler using the AWS CLI, set the following configuration option:

```
aws glue update-crawler \
   --name myCrawler \
   --configuration '{"Version": 1.0, "Grouping": {"TableGroupingPolicy":
 "CombineCompatibleSchemas" }}'
```

API

When you configure the crawler using the API, set the following configuration option:

Set the `Configuration` field with a string representation of the following JSON object in the crawler API; for example:

```
{
   "Version": 1.0,
   "Grouping": {
      "TableGroupingPolicy": "CombineCompatibleSchemas" }
}
```

Specifying the table location and partitioning level

By default, when a crawler defines tables for data stored in Amazon S3 the crawler attempts to merge schemas together, and create top-level tables (`year=2019`). In some cases, you may expect the crawler to create a table for the folder `month=Jan` but instead the crawler creates a partition since a sibling folder (`month=Mar`) was merged into the same table.

The table level crawler option provides you the flexibility to tell the crawler where the tables are located, and how you want partitions created. When you specify a **Table level**, the table is created at that absolute level from the Amazon S3 bucket.

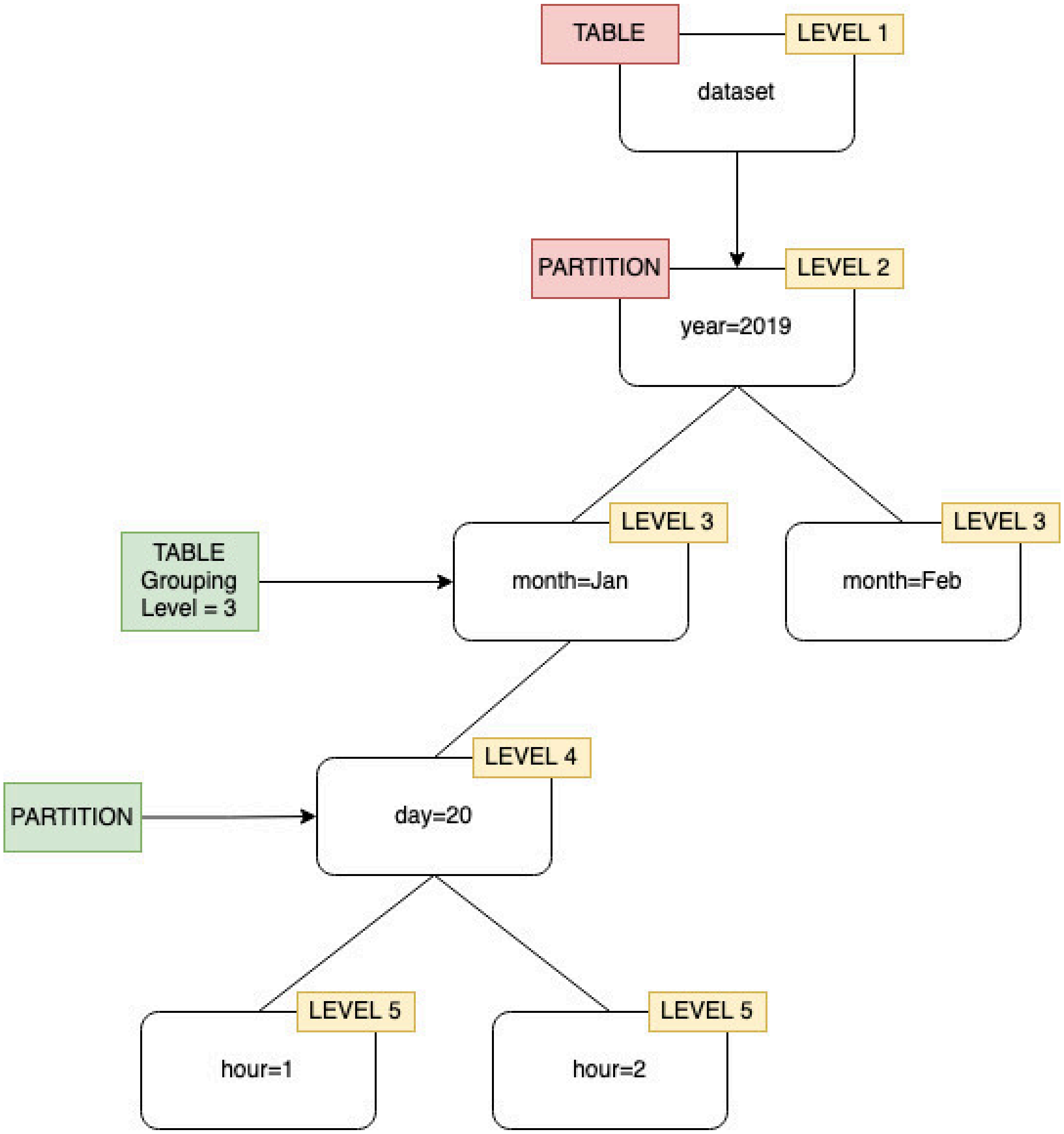


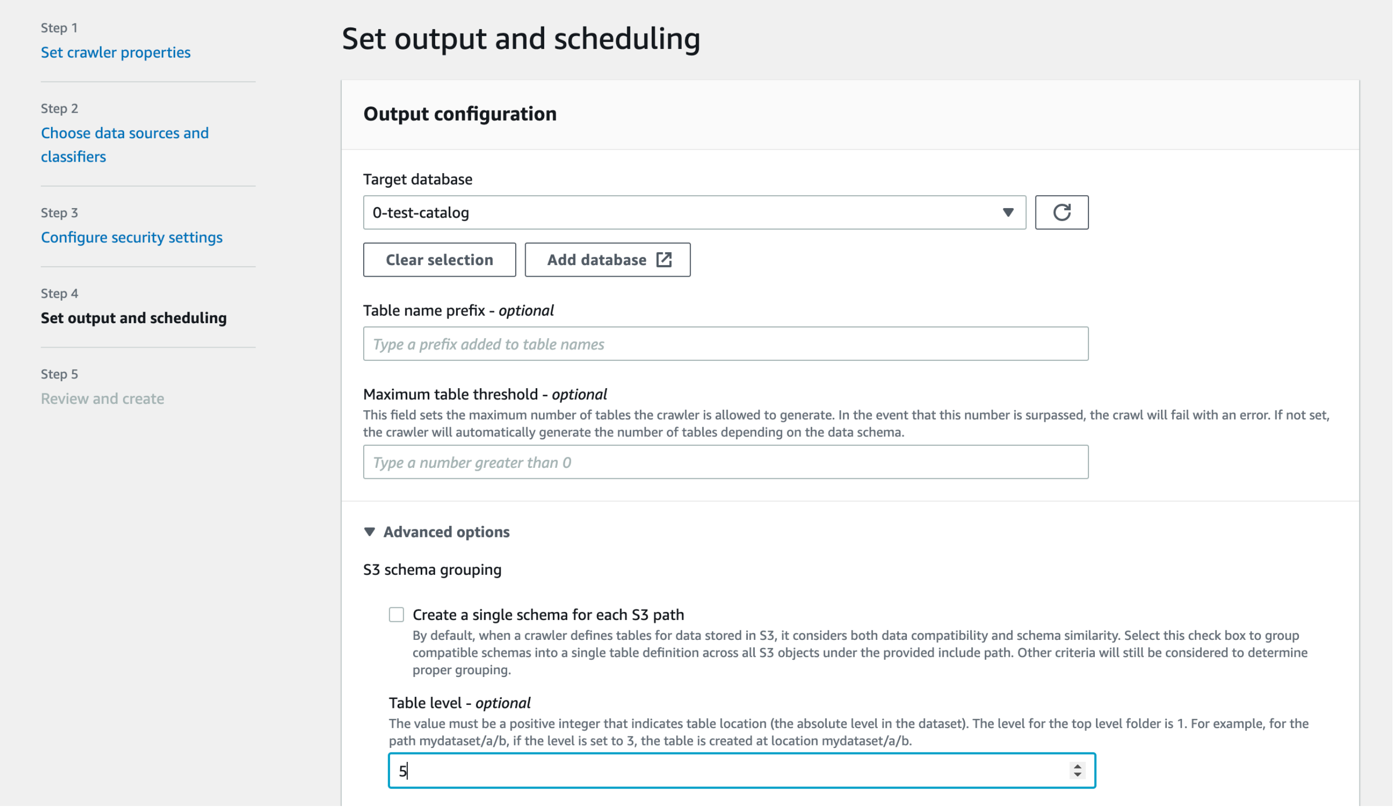
TABLE
LEVEL 1
dataset
PARTITION
LEVEL 2
year=2019
TABLE Grouping Level = 3
LEVEL 3
month=Jan
LEVEL 3
month=Feb
PARTITION
LEVEL 4
day=20
LEVEL 5
hour=1
LEVEL 5
hour=2

When configuring the crawler on the console, you can specify a value for the **Table level** crawler option. The value must be a positive integer that indicates the table location (the absolute level in the dataset). The level for the top level folder is 1. For example, for the path `mydataset/year/month/day/hour`, if the level is set to 3, the table is created at location `mydataset/year/month`.

AWS Management Console

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. Choose **Crawlers** under the **Data Catalog**.

3. When you configure a crawler, under **Output and scheduling**, choose **Table level** under **Advance options**.



Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
Clear selection
Add database

Set output and scheduling

Output configuration

Target database

0-test-catalog

Table name prefix - *optional*

Type a prefix added to table names

Maximum table threshold - *optional*

This field sets the maximum number of tables the crawler is allowed to generate. In the event that this number is surpassed, the crawl will fail with an error. If not set, the crawler will automatically generate the number of tables depending on the data schema.

Type a number greater than 0

Advanced options

S3 schema grouping

☐ Create a single schema for each S3 path

By default, when a crawler defines tables for data stored in S3, it considers both data compatibility and schema similarity. Select this check box to group compatible schemas into a single table definition across all S3 objects under the provided include path. Other criteria will still be considered to determine proper grouping.

Table level - *optional*

The value must be a positive integer that indicates table location (the absolute level in the dataset). The level for the top level folder is 1. For example, for the path mydataset/a/b, if the level is set to 3, the table is created at location mydataset/a/b.

5

AWS CLI

When you configure the crawler using the AWS CLI, set the `configuration` parameter as shown in the example code:

```
aws glue update-crawler \
  --name myCrawler \
  --configuration '{"Version": 1.0, "Grouping": { "TableLevelConfiguration": 2 }}'
```

API

When you configure the crawler using the API, set the `Configuration` field with a string representation of the following JSON object; for example:

```
configuration = jsonencode(
{
   "Version": 1.0,
    "Grouping": {
             TableLevelConfiguration = 2
         }
})
```

CloudFormation

In this example, you set the **Table level** option available in the console within your CloudFormation template:

```
"Configuration": "{
    \"Version\":1.0,
    \"Grouping\":{\"TableLevelConfiguration\":2}
}"
```

Specifying the maximum number of tables the crawler is allowed to create

You can optionally specify the maximum number of tables the crawler is allowed to create by specifying a `TableThreshold` via the AWS Glue console or AWS CLI. If the tables detected by the crawler during its crawl is greater that this input value, the crawl fails and no data is written to the Data Catalog.

This parameter is useful when the tables that would be detected and created by the crawler are much greater more than what you expect. There can be multiple reasons for this, such as:

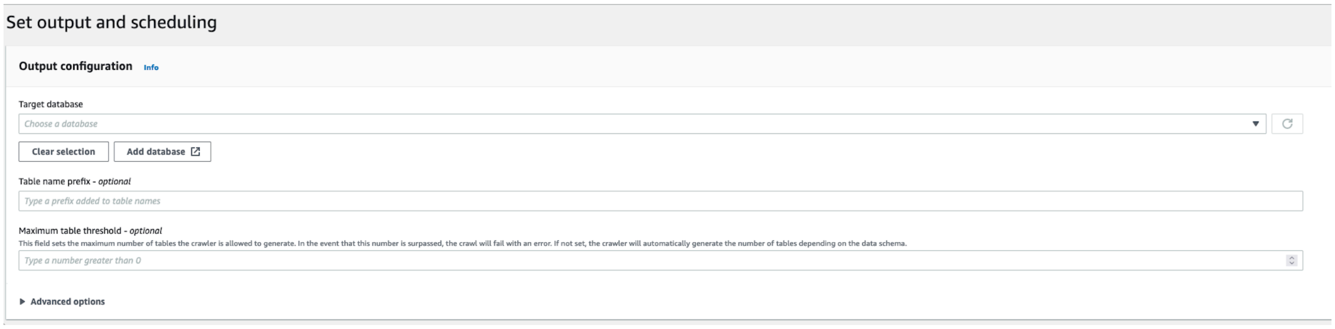
- When using an AWS Glue job to populate your Amazon S3 locations you can end up with empty files at the same level as a folder. In such cases when you run a crawler on this Amazon S3 location, the crawler creates multiple tables due to files and folders present at the same level.
- If you do not configure `"TableGroupingPolicy": "CombineCompatibleSchemas"` you may end up with more tables than expected.

You specify the `TableThreshold` as an integer value greater than 0. This value is configured on a per crawler basis. That is, for every crawl this value is considered. For example: a crawler has the `TableThreshold` value set as 5. In each crawl AWS Glue compares the number of tables detected with this table threshold value (5) and if the number of tables detected is less than 5, AWS Glue writes the tables to the Data Catalog and if not, the crawl fails without writing to the Data Catalog.

AWS Management Console

To set `TableThreshold` using the AWS Management Console:

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. When configuring a crawler, in **Output and scheduling**, set the **Maximum table threshold** to the number of tables the crawler is allowed generate.


Set output and scheduling
Output configuration
Info
Target database
Choose a database
Clear selection
Add database
Table name prefix - optional
Type a prefix added to table names
Maximum table threshold - optional
Type a number greater than 0
Advanced options

This field sets the maximum number of tables the crawler is allowed to generate. In the event that this number is surpassed, the crawl will fail with an error. If not set, the crawler will automatically generate the number of tables depending on the data schema.

AWS CLI

To set `TableThreshold` using the AWS CLI:

```
aws glue update-crawler \
    --name myCrawler \
    --configuration '{"Version": 1.0, "CrawlerOutput": {"Tables":
 { "TableThreshold": 5 }}}'
```

API

To set `TableThreshold` using the API:

```
"{"Version":1.0,
"CrawlerOutput":
{"Tables":{"AddOrUpdateBehavior":"MergeNewColumns",
"TableThreshold":5}}}";
```

Error messages are logged to help you identify table paths and clean-up your data. Example log in your account if the crawler fails because the table count was greater than table threshold value provided:

```
Table Threshold value = 28, Tables detected - 29
```

In CloudWatch, we log all table locations detected as an INFO message. An error is logged as the reason for the failure.

```
ERROR com.amazonaws.services.glue.customerLogs.CustomerLogService - CustomerLogService
 received CustomerFacingException with message
The number of tables detected by crawler: 29 is greater than the table threshold value
 provided: 28. Failing crawler without writing to Data Catalog.
com.amazonaws.services.glue.exceptions.CustomerFacingInternalException: The number of
 tables detected by crawler: 29 is greater than the table threshold value provided:
 28.
Failing crawler without writing to Data Catalog.
```

Configuring a crawler to use Lake Formation credentials

You can configure a crawler to use AWS Lake Formation credentials to access an Amazon S3 data store or a Data Catalog table with an underlying Amazon S3 location within the same AWS account or another AWS account. You can configure an existing Data Catalog table as a crawler's target, if the crawler and the Data Catalog table reside in the same account. Currently, only a single catalog target with a single catalog table is allowed when using a Data Catalog table as a crawler’s target.

> **Note**
>
> When you are defining a Data Catalog table as a crawler target, make sure that the underlying location of the Data Catalog table is an Amazon S3 location. Crawlers that use Lake Formation credentials only support Data Catalog targets with underlying Amazon S3 locations.

Setup required when the crawler and registered Amazon S3 location or Data Catalog table reside in the same account (in-account crawling)

To allow the crawler to access a data store or Data Catalog table by using Lake Formation credentials, you need to register the data location with Lake Formation. Also, the crawler's IAM role must have permissions to read the data from the destination where the Amazon S3 bucket is registered.

You can complete the following configuration steps using the AWS Management Console or AWS Command Line Interface (AWS CLI).

AWS Management Console

1. Before configuring a crawler to access the crawler source, register the data location of the data store or the Data Catalog with Lake Formation. In the Lake Formation console (https://console.aws.amazon.com/lakeformation/), register an Amazon S3 location as the root location of your data lake in the AWS account where the crawler is defined. For more information, see Registering an Amazon S3 location.
2. Grant **Data location** permissions to the IAM role that's used for the crawler run so that the crawler can read the data from the destination in Lake Formation. For more information, see Granting data location permissions (same account).
3. Grant the crawler role access permissions (`Create`) to the database, which is specified as the output database. For more information, see Granting database permissions using the Lake Formation console and the named resource method.
4. In the IAM console (https://console.aws.amazon.com/iam/), create an IAM role for the crawler. Add the `lakeformation:GetDataAccess` policy to the role.
5. In the AWS Glue console (https://console.aws.amazon.com/glue/), while configuring the crawler, select the option **Use Lake Formation credentials for crawling Amazon S3 data source**.

> **Note**
>
> The accountId field is optional for in-account crawling.

AWS CLI

```
aws glue --profile demo create-crawler --debug --cli-input-json '{
    "Name": "prod-test-crawler",
    "Role": "arn:aws:iam::111122223333:role/service-role/AWSGlueServiceRole-prod-
test-run-role",
    "DatabaseName": "prod-run-db",
    "Description": "",
    "Targets": {
    "S3Targets":[
                {
                 "Path": "s3://amzn-s3-demo-bucket"
                }
                ]
                },
    "SchemaChangePolicy": {
        "UpdateBehavior": "LOG",
        "DeleteBehavior": "LOG"
    },
    "RecrawlPolicy": {
      "RecrawlBehavior": "CRAWL_EVERYTHING"
    },
    "LineageConfiguration": {
      "CrawlerLineageSettings": "DISABLE"
    },
    "LakeFormationConfiguration": {
      "UseLakeFormationCredentials": true,
      "AccountId": "111122223333"
    },
    "Configuration": {
              "Version": 1.0,
              "CrawlerOutput": {
                "Partitions": { "AddOrUpdateBehavior": "InheritFromTable" },
                "Tables": {"AddOrUpdateBehavior": "MergeNewColumns" }
              },
              "Grouping": { "TableGroupingPolicy": "CombineCompatibleSchemas" }
            },
    "CrawlerSecurityConfiguration": "",
    "Tags": {
      "KeyName": ""
    }
  }'
```

Setup required when the crawler and registered Amazon S3 location reside in different accounts (cross-account crawling)

To allow the crawler to access a data store in a different account using Lake Formation credentials, you must first register the Amazon S3 data location with Lake Formation. Then, you grant data location permissions to the crawler's account by taking the following steps.

You can complete the following steps using the AWS Management Console or AWS CLI.

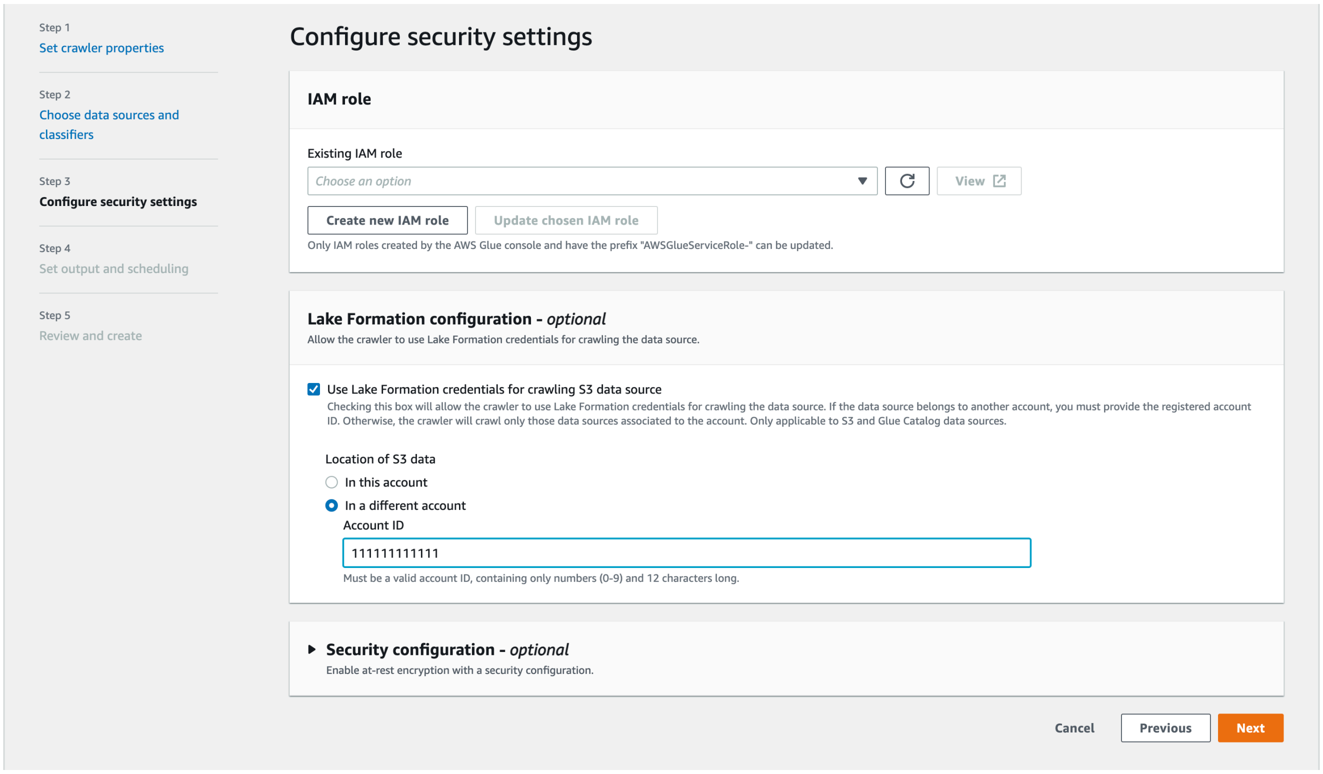
AWS Management Console

1. In the account where the Amazon S3 location is registered (account B):
   a. Register an Amazon S3 path with Lake Formation. For more information, see Registering Amazon S3 location.

    b. Grant **Data location** permissions to the account (account A) where the crawler will be run. For more information, see Grant data location permissions.
    c. Create an empty database in Lake Formation with the underlying location as the target Amazon S3 location. For more information, see Creating a database.
    d. Grant account A (the account where the crawler will be run) access to the database that you created in the previous step. For more information, see Granting database permissions.

2. In the account where the crawler is created and will be run (account A):
    a. Using the AWS RAM console, accept the database that was shared from the external account (account B). For more information, see Accepting a resource share invitation from AWS Resource Access Manager.
    b. Create an IAM role for the crawler. Add `lakeformation:GetDataAccess` policy to the role.
    c. In the Lake Formation console (https://console.aws.amazon.com/lakeformation/), grant **Data location** permissions on the target Amazon S3 location to the IAM role used for the crawler run so that the crawler can read the data from the destination in Lake Formation. For more information, see Granting data location permissions.
    d. Create a resource link on the shared database. For more information, see Create a resource link.
    e. Grant the crawler role access permissions (`Create`) on the shared database and (`Describe`) the resource link. The resource link is specified in the output for the crawler.
    f. In the AWS Glue console (https://console.aws.amazon.com/glue/), while configuring the crawler, select the option **Use Lake Formation credentials for crawling Amazon S3 data source**.

       For cross-account crawling, specify the AWS account ID where the target Amazon S3 location is registered with Lake Formation. For in-account crawling, the accountId field is optional.



Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create

Configure security settings

IAM role

Existing IAM role

Choose an option

View
Create new IAM role
Update chosen IAM role

Only IAM roles created by the AWS Glue console and have the prefix "AWSGlueServiceRole-" can be updated.

Lake Formation configuration - *optional*

Allow the crawler to use Lake Formation credentials for crawling the data source.

☑ Use Lake Formation credentials for crawling S3 data source

Checking this box will allow the crawler to use Lake Formation credentials for crawling the data source. If the data source belongs to another account, you must provide the registered account ID. Otherwise, the crawler will crawl only those data sources associated to the account. Only applicable to S3 and Glue Catalog data sources.

Location of S3 data

○ In this account

● In a different account

Account ID

111111111111

Must be a valid account ID, containing only numbers (0-9) and 12 characters long.

Security configuration - *optional*

Enable at-rest encryption with a security configuration.

Cancel
Previous
Next

AWS CLI

```
aws glue --profile demo create-crawler --debug --cli-input-json '{
    "Name": "prod-test-crawler",
    "Role": "arn:aws:iam::111122223333:role/service-role/AWSGlueServiceRole-prod-
test-run-role",
    "DatabaseName": "prod-run-db",
    "Description": "",
    "Targets": {
    "S3Targets":[
                {
                 "Path": "s3://amzn-s3-demo-bucket"
                }
                ]
                },
   "SchemaChangePolicy": {
      "UpdateBehavior": "LOG",
      "DeleteBehavior": "LOG"
  },
  "RecrawlPolicy": {
    "RecrawlBehavior": "CRAWL_EVERYTHING"
  },
  "LineageConfiguration": {
    "CrawlerLineageSettings": "DISABLE"
  },
  "LakeFormationConfiguration": {
    "UseLakeFormationCredentials": true,
    "AccountId": "111111111111"
  },
  "Configuration": {
           "Version": 1.0,
           "CrawlerOutput": {
             "Partitions": { "AddOrUpdateBehavior": "InheritFromTable" },
             "Tables": {"AddOrUpdateBehavior": "MergeNewColumns" }
           },
           "Grouping": { "TableGroupingPolicy": "CombineCompatibleSchemas" }
         },
  "CrawlerSecurityConfiguration": "",
  "Tags": {
    "KeyName": ""
  }
}'
```

Note

- A crawler using Lake Formation credentials is only supported for Amazon S3 and Data Catalog targets.
- For targets using Lake Formation credential vending, the underlying Amazon S3 locations must belong to the same bucket. For example, customers can use multiple targets (s3://amzn-s3-demo-bucket1/folder1, s3://amzn-s3-demo-bucket1/folder2) as long as all target locations are under the same bucket (amzn-s3-demo-bucket1). Specifying different buckets (s3://amzn-s3-demo-bucket1/folder1, s3://amzn-s3-demo-bucket2/folder2) is not allowed.
- Currently for Data Catalog target crawlers, only a single catalog target with a single catalog table is allowed.

Accelerating crawls using Amazon S3 event notifications

Instead of listing the objects from an Amazon S3 or Data Catalog target, you can configure the crawler to use Amazon S3 events to find any changes. This feature improves the recrawl time by using Amazon S3 events to identify the changes between two crawls by listing all the files from the subfolder which triggered the event instead of listing the full Amazon S3 or Data Catalog target.

The first crawl lists all Amazon S3 objects from the target. After the first successful crawl, you can choose to recrawl manually or on a set schedule. The crawler will list only the objects from those events instead of listing all objects.

When the target is a Data Catalog table, the crawler updates the existing tables in the Data Catalog with changes (for example, extra partitions in a table).

The advantages of moving to an Amazon S3 event based crawler are:

- A faster recrawl as the listing of all the objects from the target is not required, instead the listing of specific folders is done where objects are added or deleted.
- A reduction in the overall crawl cost as the listing of specific folders is done where objects are added or deleted.

The Amazon S3 event crawl runs by consuming Amazon S3 events from the SQS queue based on the crawler schedule. There will be no cost if there are no events in the queue. Amazon S3 events can be configured to go directly to the SQS queue or in cases where multiple consumers need the same event, a combination of SNS and SQS. For more information, see the section called “Setting up your account for Amazon S3 event notifications”.

After creating and configuring the crawler in event mode, the first crawl runs in listing mode by performing full a listing of the Amazon S3 or Data Catalog target. The following log confirms the operation of the crawl by consuming Amazon S3 events after the first successful crawl: "The crawl is running by consuming Amazon S3 events."

After creating the Amazon S3 event crawl and updating the crawler properties which may impact the crawl, the crawl operates in list mode and the following log is added: "Crawl is not running in S3 event mode".

> **Note**
>
> The maximum number of messages to consume is 100,000 messages per crawl.

Considerations and limitations

The following considerations and limitations apply when you configure a crawler to use Amazon S3 event notifications to find any changes.

- **Important behavior with deleted partitions**

  When using Amazon S3 event crawlers with Data Catalog tables:

  - If you delete a partition using the `DeletePartition` API call, you must also delete all S3 objects under that partition, and select **All object removal events** when you configure your S3 event notifications. If deletion events are not configured, the crawler recreates the deleted partition during its next run.
- Only a single target is supported by the crawler, whether for Amazon S3 or Data Catalog targets.
- SQS on private VPC is not supported.
- Amazon S3 sampling is not supported.
- The crawler target should be a folder for an Amazon S3 target, or one or more AWS Glue Data Catalog tables for a Data Catalog target.
- The 'everything' path wildcard is not supported: s3://%
- For a Data Catalog target, all catalog tables should point to same Amazon S3 bucket for Amazon S3 event mode.
- For a Data Catalog target, a catalog table should not point to an Amazon S3 location in the Delta Lake format (containing _symlink folders, or checking the catalog table's `InputFormat`).

Topics

- Setting up your account for Amazon S3 event notifications
- Setting up a crawler for Amazon S3 event notifications for an Amazon S3 target
- Setting up a crawler for Amazon S3 event notifications for a Data Catalog table

Setting up your account for Amazon S3 event notifications

Complete the following setup tasks. Note the values in parenthesis reference the configurable settings from the script.

1. You need to set up event notifications for your Amazon S3 bucket.

   For more information, see Amazon S3 event notifications.

2. To use the Amazon S3 event based crawler, you should enable event notification on the Amazon S3 bucket with events filtered from the prefix which is the same as the S3 target and store in SQS. You can set up SQS and event notification through the console by following the steps in Walkthrough: Configuring a bucket for notifications.
3. Add the following SQS policy to the role used by the crawler.

   JSON

   ```
   {
     "Version": "2012-10-17",
     "Statement": [
       {
         "Sid": "VisualEditor0",
         "Effect": "Allow",
         "Action": [
           "sqs:DeleteMessage",
           "sqs:GetQueueUrl",
           "sqs:ListDeadLetterSourceQueues",
           "sqs:ReceiveMessage",
           "sqs:GetQueueAttributes",
           "sqs:ListQueueTags",
           "sqs:SetQueueAttributes",
           "sqs:PurgeQueue"
         ],
         "Resource": "arn:aws:sqs:us-east-1:111122223333:cfn-sqs-queue"
       }
     ]
   }
   ```

Setting up a crawler for Amazon S3 event notifications for an Amazon S3 target

Follow these steps to set up a crawler for Amazon S3 event notifications for an Amazon S3 target using the AWS Management Console or AWS CLI.

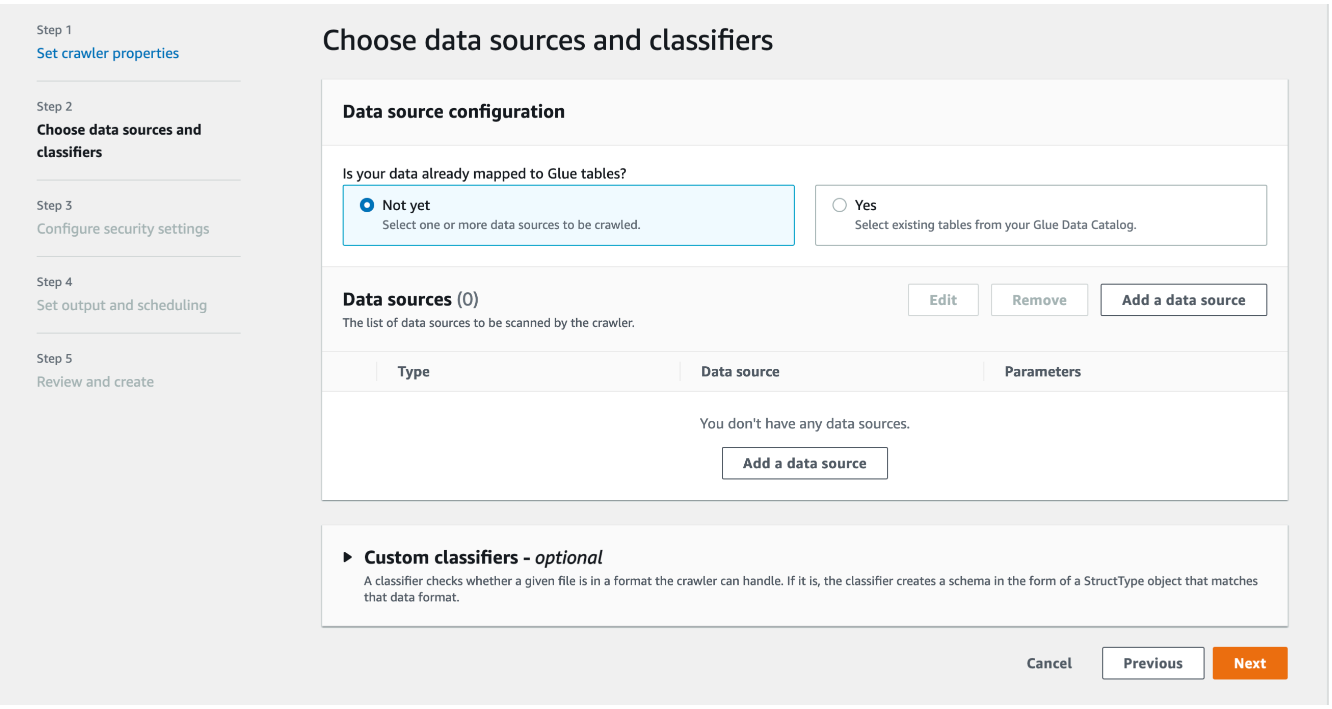
AWS Management Console

1. Sign in to the AWS Management Console and open the GuardDuty console at https://console.aws.amazon.com/guardduty/.
2. Set your crawler properties. For more information, see Setting Crawler Configuration Options on the AWS Glue console .

3. In the section **Data source configuration**, you are asked *Is your data already mapped to AWS Glue tables?*

   By default **Not yet** is already selected. Leave this as the default as you are using an Amazon S3 data source and the data is not already mapped to AWS Glue tables.

4. In the section **Data sources**, choose **Add a data source**.


Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
Edit
Remove
Add a data source
Type
Data source
Parameters
Add a data source
Cancel
Previous
Next

Choose data sources and classifiers

Data source configuration

Is your data already mapped to Glue tables?

● Not yet
Select one or more data sources to be crawled.

○ Yes
Select existing tables from your Glue Data Catalog.

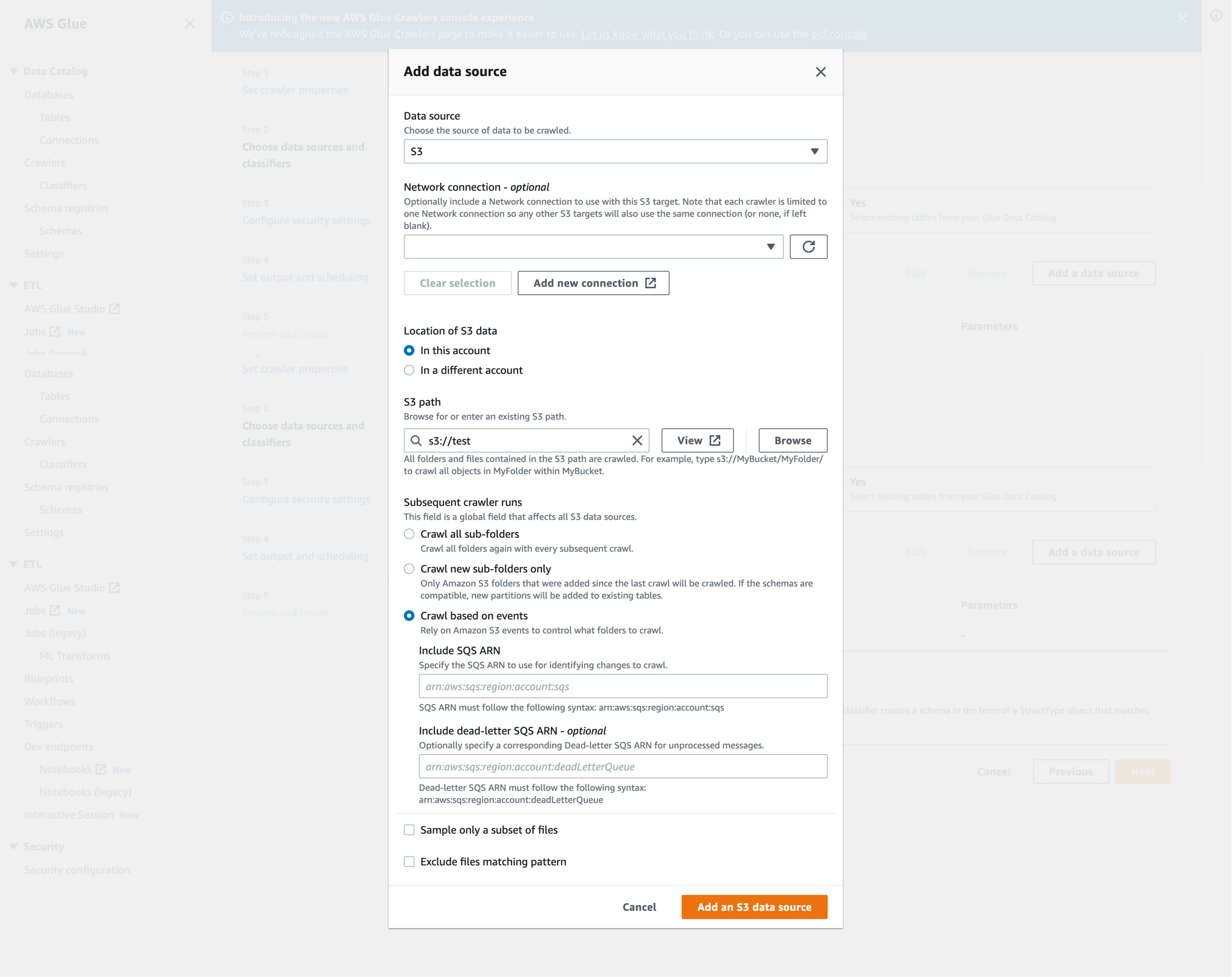
Data sources (0)
The list of data sources to be scanned by the crawler.

You don't have any data sources.

Custom classifiers - *optional*
A classifier checks whether a given file is in a format the crawler can handle. If it is, the classifier creates a schema in the form of a StructType object that matches that data format.

5. In the **Add data source** modal, configure the Amazon S3 data source:

   - **Data source**: By default, Amazon S3 is selected.
   - **Network connection** (Optional): Choose **Add new connection**.
   - **Location of Amazon S3 data**: By default, **In this account** is selected.
   - **Amazon S3 path**: Specify the Amazon S3 path where folders and files are crawled.
   - **Subsequent crawler runs**: Choose **Crawl based on events** to use Amazon S3 event notifications for your crawler.
   - **Include SQS ARN**: Specify the data store parameters including the a valid SQS ARN. (For example, `arn:aws:sqs:region:account:sqs`).
   - **Include dead-letter SQS ARN** (Optional): Specify a valid Amazon dead-letter SQS ARN. (For example, `arn:aws:sqs:region:account:deadLetterQueue`).
   - Choose **Add an Amazon S3 data source**.



Add data source

Data source
Choose the source of data to be crawled.
S3

Network connection - *optional*
Optionally include a Network connection to use with this S3 target. Note that each crawler is limited to one Network connection so any other S3 targets will also use the same connection (or none, if left blank).

Location of S3 data
● In this account
○ In a different account

S3 path
Browse for or enter an existing S3 path.
s3://test

All folders and files contained in the S3 path are crawled. For example, type s3://MyBucket/MyFolder/ to crawl all objects in MyFolder within MyBucket.

Subsequent crawler runs
This field is a global field that affects all S3 data sources.
○ Crawl all sub-folders
Crawl all folders again with every subsequent crawl.
○ Crawl new sub-folders only
Only Amazon S3 folders that were added since the last crawl will be crawled. If the schemas are compatible, new partitions will be added to existing tables.
● Crawl based on events
Rely on Amazon S3 events to control what folders to crawl.

Include SQS ARN
Specify the SQS ARN to use for identifying changes to crawl.
arn:aws:sqs:region:account:sqs
SQS ARN must follow the following syntax: arn:aws:sqs:region:account:sqs

Include dead-letter SQS ARN - *optional*
Optionally specify a corresponding Dead-letter SQS ARN for unprocessed messages.
arn:aws:sqs:region:account:deadLetterQueue
Dead-letter SQS ARN must follow the following syntax: arn:aws:sqs:region:account:deadLetterQueue

☐ Sample only a subset of files

☐ Exclude files matching pattern

AWS CLI

The following is an example Amazon S3 AWS CLI call to configure a crawler to use event notifications to crawl an Amazon S3 target bucket.

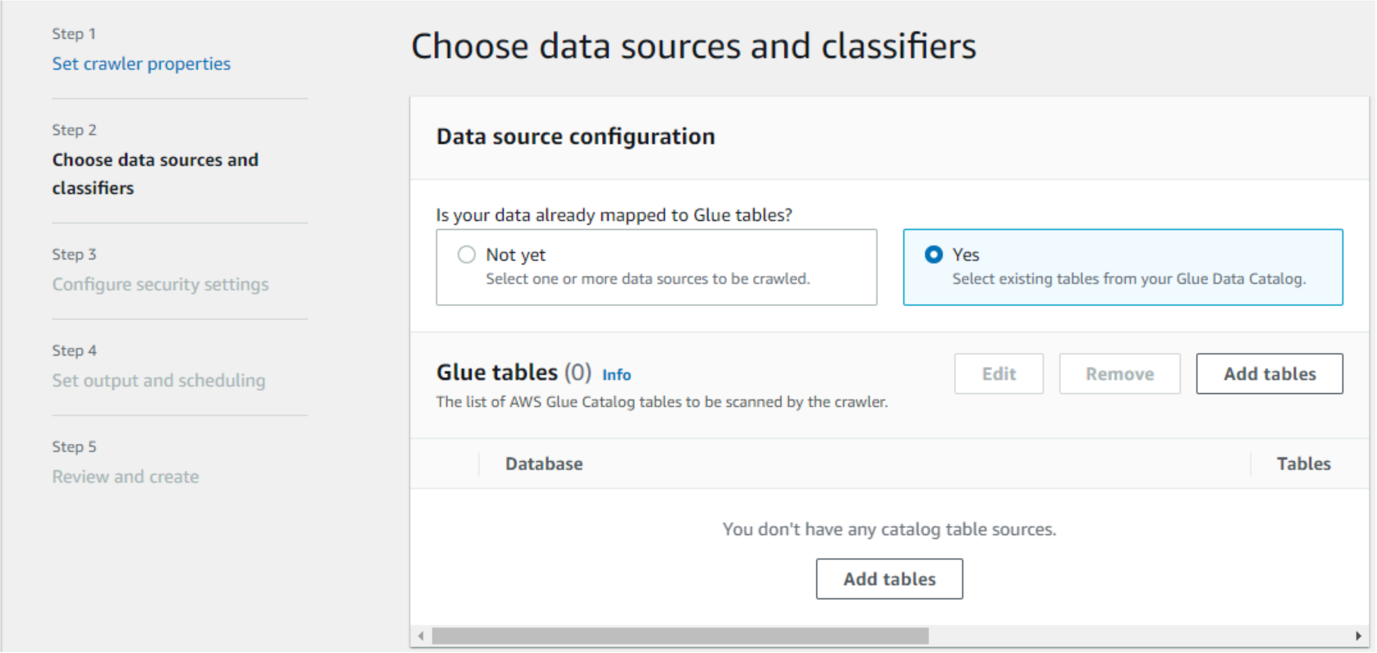
```
Create Crawler:
aws glue update-crawler \
    --name myCrawler \
    --recrawl-policy RecrawlBehavior=CRAWL_EVENT_MODE \
    --schema-change-policy UpdateBehavior=UPDATE_IN_DATABASE,DeleteBehavior=LOG
    --targets '{"S3Targets":[{"Path":"s3://amzn-s3-demo-bucket/", "EventQueueArn":
 "arn:aws:sqs:us-east-1:012345678910:MyQueue"}]}'
```

Setting up a crawler for Amazon S3 event notifications for a Data Catalog table

When you have a Data Catalog table, set up a crawler for Amazon S3 event notifications using the AWS Glue console:

1. Set your crawler properties. For more information, see Setting Crawler Configuration Options on the AWS Glue console .
2. In the section **Data source configuration**, you are asked *Is your data already mapped to AWS Glue tables?*

   Select **Yes** to select existing tables from your Data Catalog as your data source.
3. In the section **Glue tables**, choose **Add tables**.


Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
Choose data sources and classifiers
Data source configuration
Glue tables (0) Info
Edit
Remove
Add tables
Database
Tables
Add tables

Is your data already mapped to Glue tables?

○ Not yet
Select one or more data sources to be crawled.

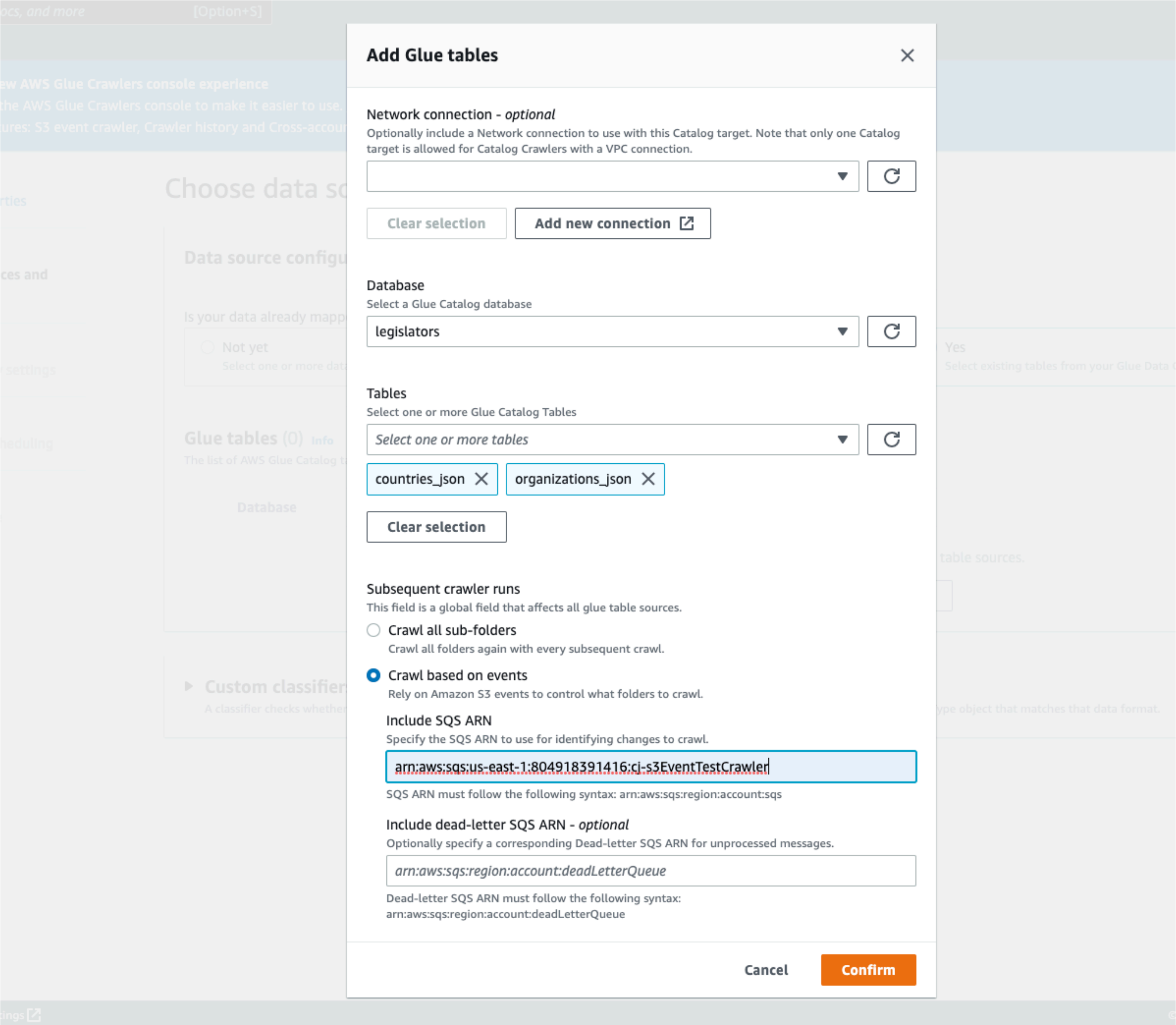
● Yes
Select existing tables from your Glue Data Catalog.

The list of AWS Glue Catalog tables to be scanned by the crawler.

You don't have any catalog table sources.

4. In the **Add table** modal, configure the database and tables:
   - **Network connection** (Optional): Choose **Add new connection**.
   - **Database**: Select a database in the Data Catalog.
   - **Tables**: Select one or more tables from that database in the Data Catalog.
   - **Subsequent crawler runs**: Choose **Crawl based on events** to use Amazon S3 event notifications for your crawler.

- **Include SQS ARN**: Specify the data store parameters including the a valid SQS ARN. (For example, `arn:aws:sqs:region:account:sqs`).
- **Include dead-letter SQS ARN** (Optional): Specify a valid Amazon dead-letter SQS ARN. (For example, `arn:aws:sqs:region:account:deadLetterQueue`).
- Choose **Confirm**.



Add Glue tables

Network connection - *optional*

Optionally include a Network connection to use with this Catalog target. Note that only one Catalog target is allowed for Catalog Crawlers with a VPC connection.

Database

Select a Glue Catalog database

legislators

Tables

Select one or more Glue Catalog Tables

countries_json

organizations_json

Subsequent crawler runs

This field is a global field that affects all glue table sources.

○ Crawl all sub-folders

Crawl all folders again with every subsequent crawl.

● Crawl based on events

Rely on Amazon S3 events to control what folders to crawl.

Include SQS ARN

Specify the SQS ARN to use for identifying changes to crawl.

arn:aws:sqs:us-east-1:804918391416:cj-s3EventTestCrawler

SQS ARN must follow the following syntax: arn:aws:sqs:region:account:sqs

Include dead-letter SQS ARN - *optional*

Optionally specify a corresponding Dead-letter SQS ARN for unprocessed messages.

Dead-letter SQS ARN must follow the following syntax: arn:aws:sqs:region:account:deadLetterQueue

Clear selection
Add new connection
Select one or more tables
Clear selection
arn:aws:sqs:region:account:deadLetterQueue
Cancel
Confirm

Tutorial: Adding an AWS Glue crawler

For this AWS Glue scenario, you're asked to analyze arrival data for major air carriers to calculate the popularity of departure airports month over month. You have flights data for the year 2016 in CSV format stored in Amazon S3. Before you transform and analyze your data, you catalog its metadata in the AWS Glue Data Catalog.

In this tutorial, let’s add a crawler that infers metadata from these flight logs in Amazon S3 and creates a table in your Data Catalog.

Topics

- Prerequisites
- Step 1: Add a crawler
- Step 2: Run the crawler
- Step 3: View AWS Glue Data Catalog objects

Prerequisites

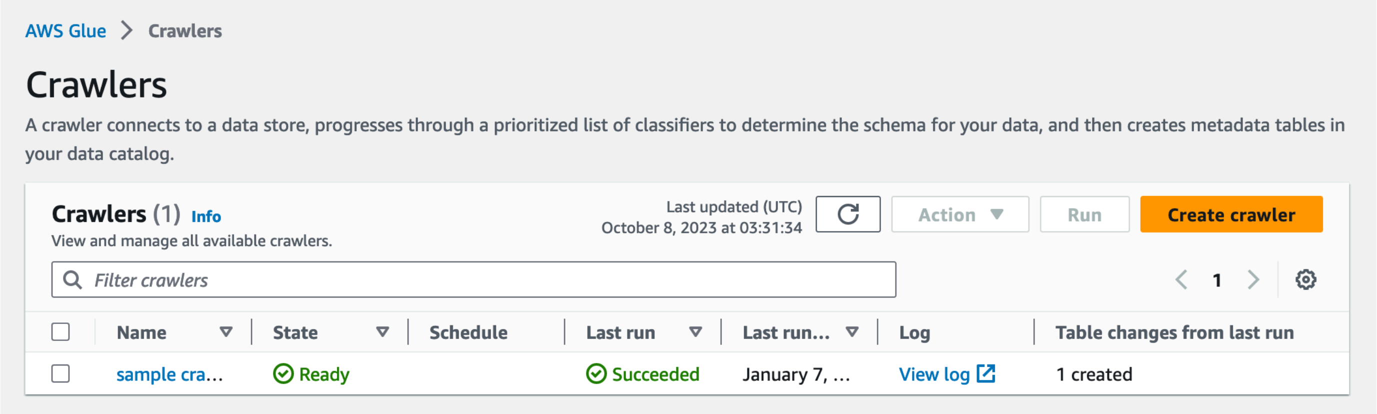
This tutorial assumes that you have an AWS account and access to AWS Glue.

Step 1: Add a crawler

Use these steps to configure and run a crawler that extracts the metadata from a CSV file stored in Amazon S3.

To create a crawler that reads files stored on Amazon S3

1. On the AWS Glue service console, on the left-side menu, choose **Crawlers**.
2. On the Crawlers page, choose **Create crawler**. This starts a series of pages that prompt you for the crawler details.



AWS Glue > Crawlers

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawlers (1) Info

View and manage all available crawlers.

Last updated (UTC)
October 8, 2023 at 03:31:34

☐	Name	State	Schedule	Last run	Last run...	Log	Table changes from last run
☐	sample cra...	Ready		Succeeded	January 7, ...	View log	1 created

3. In the Crawler name field, enter **`Flights Data Crawler`**, and choose **Next**.

   Crawlers invoke classifiers to infer the schema of your data. This tutorial uses the built-in classifier for CSV by default.

4. For the crawler source type, choose **Data stores** and choose **Next**.

5. Now let's point the crawler to your data. On the **Add a data store** page, choose the Amazon S3 data store. This tutorial doesn't use a connection, so leave the **Connection** field blank if it's visible.

   For the option **Crawl data in**, choose **Specified path in another account**. Then, for the **Include path**, enter the path where the crawler can find the flights data, which is **`s3://crawler-public-us-east-1/flight/2016/csv`**. After you enter the path, the title of this field changes to **Include path.** Choose **Next**.

6. You can crawl multiple data stores with a single crawler. However, in this tutorial, we're using only a single data store, so choose **No**, and then choose **Next**.

7. The crawler needs permissions to access the data store and create objects in the AWS Glue Data Catalog. To configure these permissions, choose **Create an IAM role**. The IAM role name starts with `AWSGlueServiceRole-`, and in the field, you enter the last part of the role name. Enter **`CrawlerTutorial`**, and then choose **Next**.

   > **Note**
   >
   > To create an IAM role, your AWS user must have `CreateRole`, `CreatePolicy`, and `AttachRolePolicy` permissions.

The wizard creates an IAM role named `AWSGlueServiceRole-CrawlerTutorial`, attaches the AWS managed policy `AWSGlueServiceRole` to this role, and adds an inline policy that allows read access to the Amazon S3 location `s3://crawler-public-us-east-1/flight/2016/csv`.

8. Create a schedule for the crawler. For **Frequency**, choose **Run on demand**, and then choose **Next**.
9. Crawlers create tables in your Data Catalog. Tables are contained in a database in the Data Catalog. First, choose **Add database** to create a database. In the pop-up window, enter **`test-flights-db`** for the database name, and then choose **Create**.

   Next, enter **`flights`** for **Prefix added to tables**. Use the default values for the rest of the options, and choose **Next**.
10. Verify the choices you made in the **Add crawler** wizard. If you see any mistakes, you can choose **Back** to return to previous pages and make changes.

    After you have reviewed the information, choose **Finish** to create the crawler.

Step 2: Run the crawler

After creating a crawler, the wizard sends you to the Crawlers view page. Because you create the crawler with an on-demand schedule, you're given the option to run the crawler.

To run the crawler

1. The banner near the top of this page lets you know that the crawler was created, and asks if you want to run it now. Choose **Run it now?** to run the crawler.

   The banner changes to show "Attempting to run" and Running" messages for your crawler. After the crawler starts running, the banner disappears, and the crawler display is updated to show a status of Starting for your crawler. After a minute, you can click the Refresh icon to update the status of the crawler that is displayed in the table.
2. When the crawler completes, a new banner appears that describes the changes made by the crawler. You can choose the **test-flights-db** link to view the Data Catalog objects.

Step 3: View AWS Glue Data Catalog objects

The crawler reads data at the source location and creates tables in the Data Catalog. A table is the metadata definition that represents your data, including its schema. The tables in the Data Catalog do not contain data. Instead, you use these tables as a source or target in a job definition.

To view the Data Catalog objects created by the crawler

1. In the left-side navigation, under **Data catalog**, choose **Databases**. Here you can view the `flights-db` database that is created by the crawler.
2. In the left-side navigation, under **Data catalog** and below **Databases**, choose **Tables**. Here you can view the `flightscsv` table created by the crawler. If you choose the table name, then you can view the table settings, parameters, and properties. Scrolling down in this view, you can view the schema, which is information about the columns and data types of the table.
3. If you choose **View partitions** on the table view page, you can see the partitions created for the data. The first column is the partition key.

Defining metadata manually

The AWS Glue Data Catalog is a central repository that stores metadata about your data sources and data sets. While a crawler can automatically crawl and populate metadata for supported data sources, there are certain scenarios where you may need to define metadata manually in the Data Catalog:

- Unsupported data formats – If you have data sources that are not supported by the crawler, you need to manually define the metadata for those data sources in the Data Catalog.
- Custom metadata requirements – The AWS Glue crawler infers metadata based on predefined rules and conventions. If you have specific metadata requirements that are not covered by the AWS Glue crawler inferred metadata, you can manually define the metadata to meet your needs
- Data governance and standardization – In some cases, you may want to have more control over the metadata definitions for data governance, compliance, or security reasons. Manually defining metadata allows you to ensure that the metadata adheres to your organization's standards and policies.
- Placeholder for future data ingestion – If you have data sources that are not immediately available or accessible, you can create empty schema tables as placeholders. Once the data sources become available, you can populate the tables with the actual data, while maintaining the predefined structure.

To define metadata manually, you can use the AWS Glue console, Lake Formation console, AWS Glue API, or the AWS Command Line Interface (AWS CLI). You can create databases, tables, and partitions, and specify metadata properties such as column names, data types, descriptions, and other attributes.

Creating databases

Databases are used to organize metadata tables in the AWS Glue. When you define a table in the AWS Glue Data Catalog, you add it to a database. A table can be in only one database.

Your database can contain tables that define data from many different data stores. This data can include objects in Amazon Simple Storage Service (Amazon S3) and relational tables in Amazon Relational Database Service.

> **Note**
>
> When you delete a database from the AWS Glue Data Catalog, all the tables in the database are also deleted.

To view the list of databases, sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/. Choose **Databases**, and then choose a database name in the list to view the details.

From the **Databases** tab in the AWS Glue console, you can add, edit, and delete databases:

- To create a new database, choose **Add database** and provide a name and description. For compatibility with other metadata stores, such as Apache Hive, the name is folded to lowercase characters.

  > **Note**
  >
  > If you plan to access the database from Amazon Athena, then provide a name with only alphanumeric and underscore characters. For more information, see Athena names.

- To edit the description for a database, select the check box next to the database name and choose **Edit**.
- To delete a database, select the check box next to the database name and choose **Remove**.

- To display the list of tables contained in the database, choose the database name and the database properties will display all tables in the database.

To change the database that a crawler writes to, you must change the crawler definition. For more information, see Using crawlers to populate the Data Catalog .

Database resource links

The AWS Glue console was recently updated. The current version of the console does not support Database Resource Links.

The Data Catalog can also contain *resource links* to databases. A database resource link is a link to a local or shared database. Currently, you can create resource links only in AWS Lake Formation. After you create a resource link to a database, you can use the resource link name wherever you would use the database name. Along with databases that you own or that are shared with you, database resource links are returned by `glue:GetDatabases()` and appear as entries on the **Databases** page of the AWS Glue console.

The Data Catalog can also contain table resource links.

For more information about resource links, see Creating Resource Links in the *AWS Lake Formation Developer Guide*.

Creating tables

Even though running a crawler is the recommended method to take inventory of the data in your data stores, you can add metadata tables to the AWS Glue Data Catalog manually. This approach allows you to have more control over the metadata definitions and customize them according them to your specific requirements.

You can also add tables to the Data Catalog manually in the following ways:

- Use the AWS Glue console to manually create a table in the AWS Glue Data Catalog. For more information, see Creating tables using the console.
- Use the `CreateTable` operation in the AWS Glue API to create a table in the AWS Glue Data Catalog. For more information, see CreateTable action (Python: create_table).
- Use AWS CloudFormation templates. For more information, see AWS CloudFormation for AWS Glue.

When you define a table manually using the console or an API, you specify the table schema and the value of a classification field that indicates the type and format of the data in the data source. If a crawler creates the table, the data format and schema are determined by either a built-in classifier or a custom classifier. For more information about creating a table using the AWS Glue console, see Creating tables using the console.

Topics

- Table partitions
- Table resource links
- Creating tables using the console
- Creating partition indexes
- Updating manually created Data Catalog tables using crawlers
- Data Catalog table properties

Table partitions

An AWS Glue table definition of an Amazon Simple Storage Service (Amazon S3) folder can describe a partitioned table. For example, to improve query performance, a partitioned table might separate monthly data into different files using the name of the month as a key. In AWS Glue, table definitions include the partitioning key of a table. When AWS Glue evaluates the data in Amazon S3 folders to catalog a table, it determines whether an individual table or a partitioned table is added.

You can create partition indexes on a table to fetch a subset of the partitions instead of loading all the partitions in the table. For information about working with partition indexes, see Creating partition indexes .

All the following conditions must be true for AWS Glue to create a partitioned table for an Amazon S3 folder:

- The schemas of the files are similar, as determined by AWS Glue.
- The data format of the files is the same.
- The compression format of the files is the same.

For example, you might own an Amazon S3 bucket named `my-app-bucket`, where you store both iOS and Android app sales data. The data is partitioned by year, month, and day. The data files for iOS and Android sales have the same schema, data format, and compression format. In the AWS Glue Data Catalog, the AWS Glue crawler creates one table definition with partitioning keys for year, month, and day.

The following Amazon S3 listing of `my-app-bucket` shows some of the partitions. The = symbol is used to assign partition key values.

```
my-app-bucket/Sales/year=2010/month=feb/day=1/iOS.csv
my-app-bucket/Sales/year=2010/month=feb/day=1/Android.csv
my-app-bucket/Sales/year=2010/month=feb/day=2/iOS.csv
my-app-bucket/Sales/year=2010/month=feb/day=2/Android.csv
...
my-app-bucket/Sales/year=2017/month=feb/day=4/iOS.csv
my-app-bucket/Sales/year=2017/month=feb/day=4/Android.csv
```

Table resource links

> The AWS Glue console was recently updated. The current version of the console does not support Table Resource Links.

The Data Catalog can also contain *resource links* to tables. A table resource link is a link to a local or shared table. Currently, you can create resource links only in AWS Lake Formation. After you create a resource link to a table, you can use the resource link name wherever you would use the table name. Along with tables that you own or that are shared with you, table resource links are returned by `glue:GetTables()` and appear as entries on the **Tables** page of the AWS Glue console.

The Data Catalog can also contain database resource links.

For more information about resource links, see Creating Resource Links in the *AWS Lake Formation Developer Guide*.

Creating tables using the console

A table in the AWS Glue Data Catalog is the metadata definition that represents the data in a data store. You create tables when you run a crawler, or you can create a table manually in the AWS Glue console. The **Tables** list in the AWS Glue console displays values of your table's metadata. You use table definitions to specify sources and targets when you create ETL (extract, transform, and load) jobs.

Note

With recent changes to the AWS management console, you may need to modify your existing IAM roles to have the `SearchTables` permission. For new role creation, the `SearchTables` API permission has already been added as default.

To get started, sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/. Choose the **Tables** tab, and use the **Add tables** button to create tables either with a crawler or by manually typing attributes.

Adding tables on the console

To use a crawler to add tables, choose **Add tables**, **Add tables using a crawler**. Then follow the instructions in the **Add crawler** wizard. When the crawler runs, tables are added to the AWS Glue Data Catalog. For more information, see Using crawlers to populate the Data Catalog .

If you know the attributes that are required to create an Amazon Simple Storage Service (Amazon S3) table definition in your Data Catalog, you can create it with the table wizard. Choose **Add tables**, **Add table manually**, and follow the instructions in the **Add table** wizard.

When adding a table manually through the console, consider the following:

- If you plan to access the table from Amazon Athena, then provide a name with only alphanumeric and underscore characters. For more information, see Athena names.
- The location of your source data must be an Amazon S3 path.
- The data format of the data must match one of the listed formats in the wizard. The corresponding classification, SerDe, and other table properties are automatically populated based on the format chosen. You can define tables with the following formats:

  Avro

  Apache Avro JSON binary format.

  CSV

  Character separated values. You also specify the delimiter of either comma, pipe, semicolon, tab, or Ctrl-A.

  JSON

  JavaScript Object Notation.

XML

Extensible Markup Language format. Specify the XML tag that defines a row in the data. Columns are defined within row tags.

Parquet

Apache Parquet columnar storage.

ORC

Optimized Row Columnar (ORC) file format. A format designed to efficiently store Hive data.

- You can define a partition key for the table.
- Currently, partitioned tables that you create with the console cannot be used in ETL jobs.

Table attributes

The following are some important attributes of your table:

Name

The name is determined when the table is created, and you can't change it. You refer to a table name in many AWS Glue operations.

Database

The container object where your table resides. This object contains an organization of your tables that exists within the AWS Glue Data Catalog and might differ from an organization in your data store. When you delete a database, all tables contained in the database are also deleted from the Data Catalog.

Description

The description of the table. You can write a description to help you understand the contents of the table.

Table format

Specify creating a standard AWS Glue table, or a table in Apache Iceberg format.

The Data Catalog provides following table optimization options to manage table storage and improve query performance for Iceberg tables.

- **Compaction –** Data files are merged and rewritten remove obsolete data and consolidate fragmented data into larger, more efficient files.
- **Snapshot retention –** Snapshots are timestamped versions of an Iceberg table. Snapshot retention configurations allow customers to enforce how long to retain snapshots and how many snapshots to retain. Configuring a snapshot retention optimizer can help manage storage overhead by removing older, unnecessary snapshots and their associated underlying files.
- **Orphan file deletion –** Orphan files are files that are no longer referenced by the Iceberg table metadata. These files can accumulate over time, especially after operations like table deletions or failed ETL jobs. Enabling orphan file deletion allows AWS Glue to periodically identify and remove these unnecessary files, freeing up storage.

For more information, see Optimizing Iceberg tables.

Optimization configuration

You can either use the default settings or customize the settings for enabling the table optimizers.

IAM role

To run the table optimizers, the service assumes an IAM role on your behalf. You can choose an IAM role using the drop-down. Ensure that the role has the permissions required to enable compaction.

To learn more about the required permissions for the IAM role, see Table optimization prerequisites .

Location

The pointer to the location of the data in a data store that this table definition represents.

Classification

A categorization value provided when the table was created. Typically, this is written when a crawler runs and specifies the format of the source data.

Last updated

The time and date (UTC) that this table was updated in the Data Catalog.

Date added

The time and date (UTC) that this table was added to the Data Catalog.

Deprecated

If AWS Glue discovers that a table in the Data Catalog no longer exists in its original data store, it marks the table as deprecated in the data catalog. If you run a job that references a deprecated table, the job might fail. Edit jobs that reference deprecated tables to remove them as sources and targets. We recommend that you delete deprecated tables when they are no longer needed.

Connection

If AWS Glue requires a connection to your data store, the name of the connection is associated with the table.

Viewing and managing table details

To see the details of an existing table, choose the table name in the list, and then choose **Action, View details**.

The table details include properties of your table and its schema. This view displays the schema of the table, including column names in the order defined for the table, data types, and key columns for partitions. If a column is a complex type, you can choose **View properties** to display details of the structure of that field, as shown in the following example:

```
{
"StorageDescriptor":
    {
      "cols": {
         "FieldSchema": [
           {
             "name": "primary-1",
             "type": "CHAR",
             "comment": ""
           },
           {
             "name": "second ",
             "type": "STRING",
             "comment": ""
           }
         ]
      },
      "location": "s3://aws-logs-111122223333-us-east-1",
      "inputFormat": "",
      "outputFormat": "org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat",
      "compressed": "false",
      "numBuckets": "0",
      "SerDeInfo": {
           "name": "",
           "serializationLib": "org.apache.hadoop.hive.serde2.OpenCSVSerde",
           "parameters": {
               "separatorChar": "|"
            }
      },
      "bucketCols": [],
      "sortCols": [],
      "parameters": {},
      "SkewedInfo": {},
      "storedAsSubDirectories": "false"
    },
    "parameters": {
       "classification": "csv"
    }
}
```

For more information about the properties of a table, such as `StorageDescriptor`, see StorageDescriptor structure.

To change the schema of a table, choose **Edit schema** to add and remove columns, change column names, and change data types.

To compare different versions of a table, including its schema, choose **Compare versions** to see a side-by-side comparison of two versions of the schema for a table. For more information, see Comparing table schema versions .

To display the files that make up an Amazon S3 partition, choose **View partition**. For Amazon S3 tables, the **Key** column displays the partition keys that are used to partition the table in the source data store. Partitioning is a way to divide a table into related parts based on the values of a key column, such as date, location, or department. For more information about partitions, search the internet for information about "hive partitioning."

Note

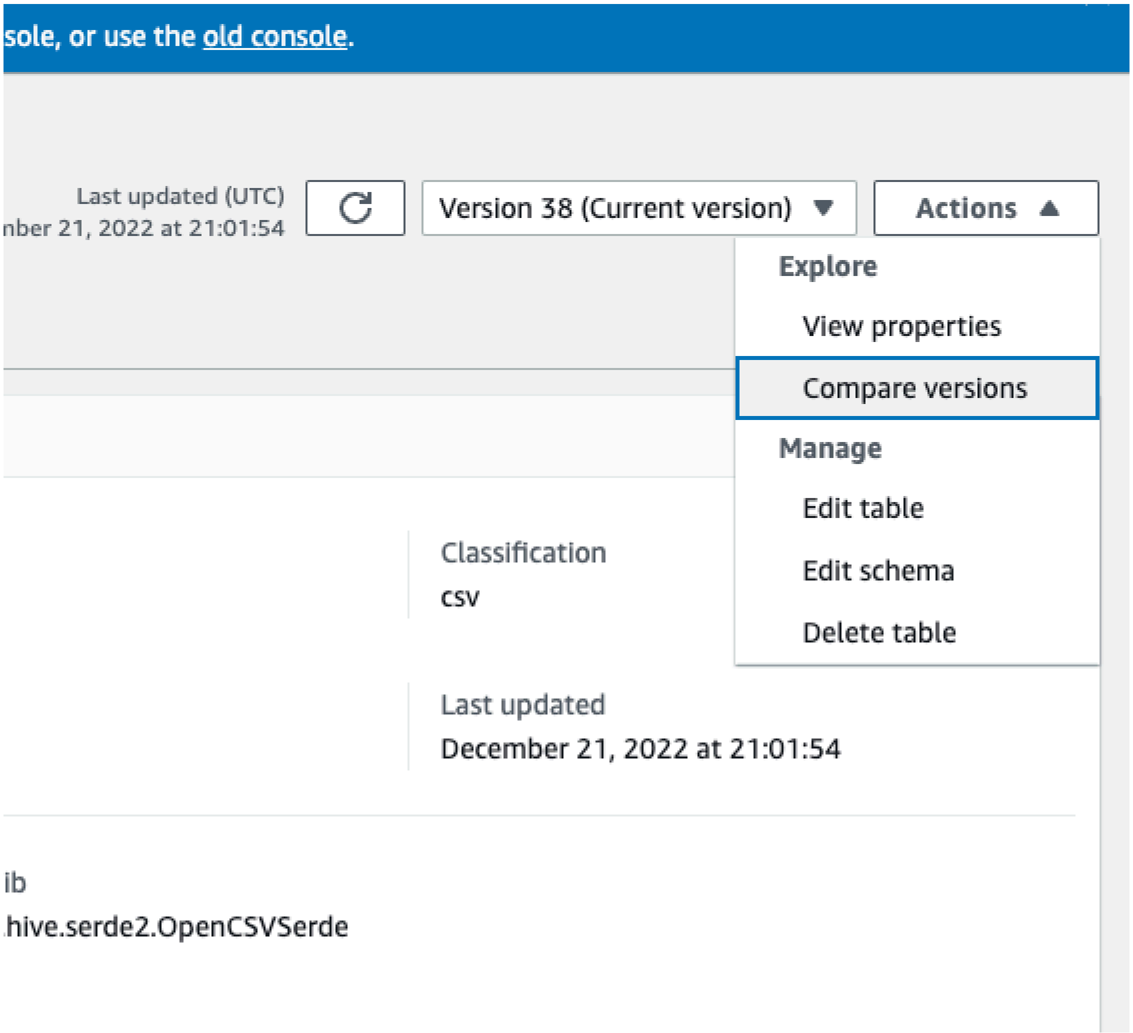
To get step-by-step guidance for viewing the details of a table, see the **Explore table** tutorial in the console.

Comparing table schema versions

When you compare two versions of table schemas, you can compare nested row changes by expanding and collapsing nested rows, compare schemas of two versions side-by-side, and view table properties side-by-side.

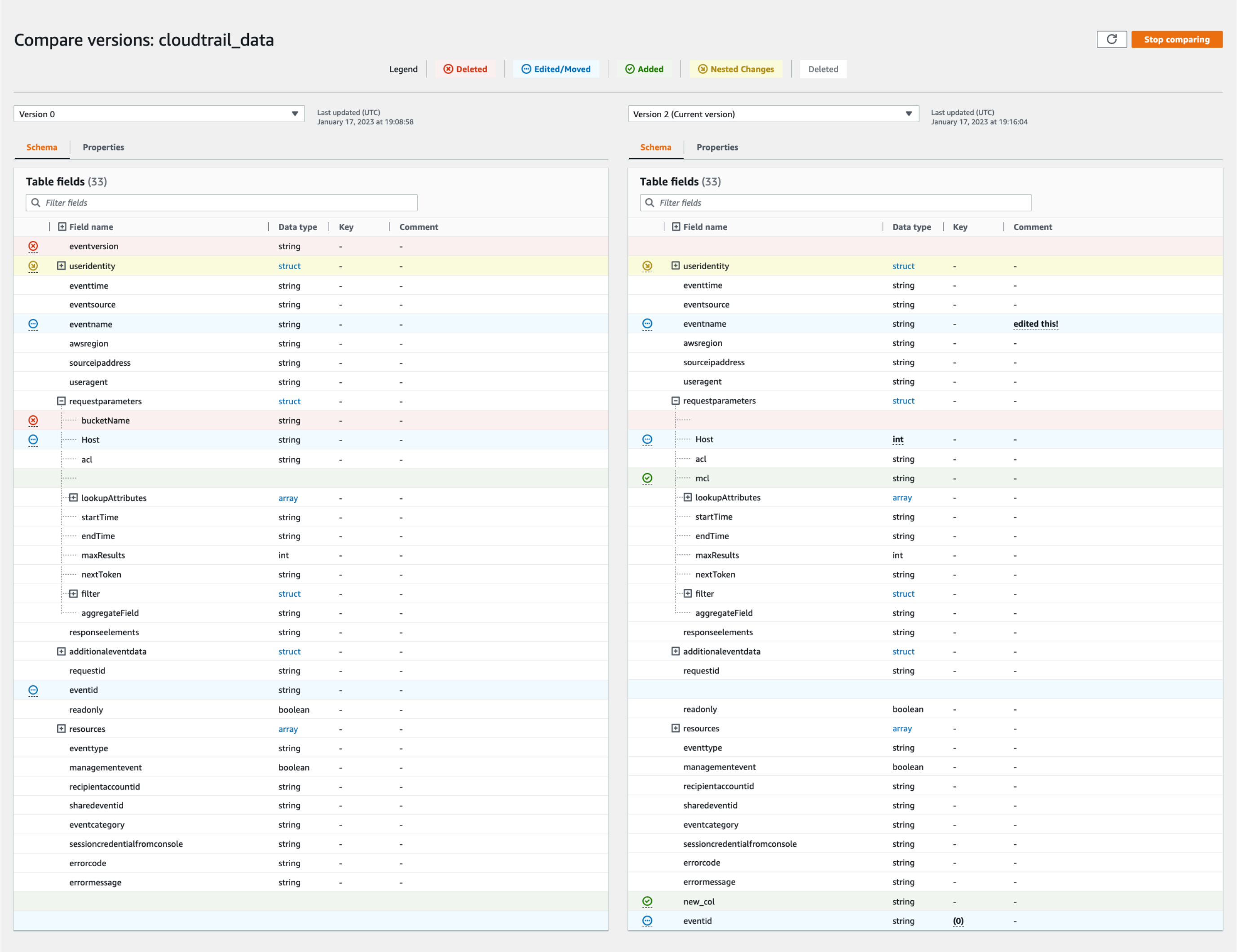
To compare versions

1. From the AWS Glue console, choose **Tables**, then **Actions** and choose **Compare versions**.


sole, or use the old console.
Last updated (UTC)
nber 21, 2022 at 21:01:54
Version 38 (Current version)
Actions
Explore
View properties
Compare versions
Manage
Edit table
Edit schema
Delete table
Classification
csv
Last updated
December 21, 2022 at 21:01:54
ib
hive.serde2.OpenCSVSerde

2. Choose a version to compare by choosing the version drop-down menu. When comparing schemas, the Schema tab is highlighted in orange.
3. When you compare tables between two versions, the table schemas are presented to you on the left and right side of the screen. This enables you to determine changes visually by comparing the Column name, data type, key, and comment fields side-by-side. When there is a change, a colored icon displays the type of change that was made.

- Deleted – displayed by a red icon indicates where the column was removed from a previous version of the table schema.
- Edited or Moved – displayed by a blue icon indicates where the column was modified or moved in a newer version of the table schema.
- Added – displayed by a green icon indicates where the column was added to a newer version of the table schema.
- Nested changes – displayed by a yellow icon indicates where the nested column contains changes. Choose the column to expand and view the columns that have either been deleted, edited, moved, or added.



Compare versions: cloudtrail_data

Legend: ⊗ Deleted | ⊖ Edited/Moved | ✓ Added | ⊙ Nested Changes | Deleted

Version 0 — Last updated (UTC) January 17, 2023 at 19:08:58

Table fields (33)

	⊞ Field name	Data type	Key	Comment
⊗	eventversion	string	-	-
⊙	⊞ useridentity	struct	-	-
	eventtime	string	-	-
	eventsource	string	-	-
⊖	eventname	string	-	-
	awsregion	string	-	-
	sourceipaddress	string	-	-
	useragent	string	-	-
	⊟ requestparameters	struct	-	-
⊗	bucketName	string	-	-
⊖	Host	string	-	-
	acl	string	-	-
	⊞ lookupAttributes	array	-	-
	startTime	string	-	-
	endTime	string	-	-
	maxResults	int	-	-
	nextToken	string	-	-
	⊞ filter	struct	-	-
	aggregateField	string	-	-
	responseelements	string	-	-
	⊞ additionaleventdata	struct	-	-
	requestid	string	-	-
⊖	eventid	string	-	-
	readonly	boolean	-	-
	⊞ resources	array	-	-
	eventtype	string	-	-
	managementevent	boolean	-	-
	recipientaccountid	string	-	-
	sharedeventid	string	-	-
	eventcategory	string	-	-
	sessioncredentialfromconsole	string	-	-
	errorcode	string	-	-
	errormessage	string	-	-

Version 2 (Current version) — Last updated (UTC) January 17, 2023 at 19:16:04

Table fields (33)

	⊞ Field name	Data type	Key	Comment
⊙	⊞ useridentity	struct	-	-
	eventtime	string	-	-
	eventsource	string	-	-
⊖	eventname	string	-	edited this!
	awsregion	string	-	-
	sourceipaddress	string	-	-
	useragent	string	-	-
	⊟ requestparameters	struct	-	-
⊖	Host	int	-	-
	acl	string	-	-
✓	mcl	string	-	-
	⊞ lookupAttributes	array	-	-
	startTime	string	-	-
	endTime	string	-	-
	maxResults	int	-	-
	nextToken	string	-	-
	⊞ filter	struct	-	-
	aggregateField	string	-	-
	responseelements	string	-	-
	⊞ additionaleventdata	struct	-	-
	requestid	string	-	-
	readonly	boolean	-	-
	⊞ resources	array	-	-
	eventtype	string	-	-
	managementevent	boolean	-	-
	recipientaccountid	string	-	-
	sharedeventid	string	-	-
	eventcategory	string	-	-
	sessioncredentialfromconsole	string	-	-
	errorcode	string	-	-
	errormessage	string	-	-
✓	new_col	string	-	-
⊖	eventid	string	(0)	-

4. Use the filter fields search bar to display fields based on the characters you enter here. If you enter a column name in either table version, the filtered fields are displayed in both table versions to show you where the changes have occurred.
5. To compare properties, choose the **Properties tab**.
6. To stop comparing versions, choose **Stop comparing** to return to the list of tables.

Creating partition indexes

Over time, hundreds of thousands of partitions get added to a table. The GetPartitions API is used to fetch the partitions in the table. The API returns partitions that match the expression provided in the request.

Lets take a *sales_data* table as an example which is partitioned by the keys *Country*, *Category*, *Year*, *Month*, and *creationDate*. If you want to obtain sales data for all the items sold for the *Books* category in the year 2020 after *2020-08-15*, you have to make a `GetPartitions` request with the expression "Category = 'Books' and creationDate > '2020-08-15'" to the Data Catalog.

If no partition indexes are present on the table, AWS Glue loads all the partitions of the table, and then filters the loaded partitions using the query expression provided by the user in the `GetPartitions` request. The query takes more time to run as the number of partitions increase on a table with no indexes. With an index, the `GetPartitions` query will try to fetch a subset of the partitions instead of loading all the partitions in the table.

Topics

- About partition indexes
- Creating a table with partition indexes
- Adding a partition index to an existing table
- Describing partition indexes on a table
- Limitations on using partition indexes
- Using indexes for an optimized GetPartitions call
- Integration with engines

About partition indexes

When you create a partition index, you specify a list of partition keys that already exist on a given table. Partition index is sub list of partition keys defined in the table. A partition index can be created on any permutation of partition keys defined on the table. For the above *sales_data* table, the possible indexes are (country, category, creationDate), (country, category, year), (country, category), (country), (category, country, year, month), and so on.

The Data Catalog will concatenate the partition values in the order provided at the time of index creation. The index is built consistently as partitions are added to the table. Indexes can be created for String (string, char, and varchar), Numeric (int, bigint, long, tinyint, and smallint), and Date (yyyy-MM-dd) column types.

Supported data types

- Date – A date in ISO format, such as `YYYY-MM-DD`. For example, date `2020-08-15`. The format uses hyphens (-) to separate the year, month, and day. The permissible range for dates for indexing spans from `0000-01-01` to `9999-12-31`.
- String – A string literal enclosed in single or double quotes.
- Char – Fixed length character data, with a specified length between 1 and 255, such as char(10).
- Varchar – Variable length character data, with a specified length between 1 and 65535, such as varchar(10).
- Numeric – int, bigint, long, tinyint, and smallint

Indexes on Numeric, String, and Date data types support =, >, >=, <, <= and between operators. The indexing solution currently only supports the `AND` logical operator. Sub-expressions with the operators "LIKE", "IN", "OR", and "NOT" are ignored in the expression for filtering using an index. Filtering for the ignored sub-expression is done on the partitions fetched after applying index filtering.

For each partition added to a table, there is a corresponding index item created. For a table with ‘n’ partitions, 1 partition index will result in 'n' partition index items. 'm' partition index on same table will result into 'm*n' partition index items. Each partition index item will be charged according to the current AWS Glue pricing policy for data catalog storage. For details on storage object pricing, see AWS Glue pricing.

Creating a table with partition indexes

You can create a partition index during table creation. The `CreateTable` request takes a list of `PartitionIndex` objects as an input. A maximum of 3 partition indexes can be created on a given table. Each partition index requires a name and a list of `partitionKeys` defined for the table. Created indexes on a table can be fetched using the `GetPartitionIndexes` API

Adding a partition index to an existing table

To add a partition index to an existing table, use the `CreatePartitionIndex` operation. You can create one `PartitionIndex` per `CreatePartitionIndex` operation. Adding an index does not affect the availability of a table, as the table continues to be available while indexes are being created.

The index status for an added partition is set to CREATING and the creation of the index data is started. If the process for creating the indexes is successful, the indexStatus is updated to ACTIVE and for an unsuccessful process, the index status is updated to FAILED. Index creation can fail for multiple reasons, and you can use the `GetPartitionIndexes` operation to retrieve the failure details. The possible failures are:

- ENCRYPTED_PARTITION_ERROR — Index creation on a table with encrypted partitions is not supported.
- INVALID_PARTITION_TYPE_DATA_ERROR — Observed when the `partitionKey` value is not a valid value for the corresponding `partitionKey` data type. For example: a `partitionKey` with the 'int' datatype has a value 'foo'.
- MISSING_PARTITION_VALUE_ERROR — Observed when the `partitionValue` for an `indexedKey` is not present. This can happen when a table is not partitioned consistently.
- UNSUPPORTED_PARTITION_CHARACTER_ERROR — Observed when the value for an indexed partition key contains the characters \u0000, \u0001 or \u0002
- INTERNAL_ERROR — An internal error occurred while indexes were being created.

Describing partition indexes on a table

To fetch the partition indexes created on a table, use the `GetPartitionIndexes` operation. The response returns all the indexes on the table, along with the current status of each index (the `IndexStatus`).

The `IndexStatus` for a partition index will be one of the following:

- `CREATING` — The index is currently being created, and is not yet available for use.
- `ACTIVE` — The index is ready for use. Requests can use the index to perform an optimized query.
- `DELETING` — The index is currently being deleted, and can no longer be used. An index in the active state can be deleted using the `DeletePartitionIndex` request, which moves the status from ACTIVE to DELETING.

- `FAILED` — The index creation on an existing table failed. Each table stores the last 10 failed indexes.

The possible state transitions for indexes created on an existing table are:

- CREATING → ACTIVE → DELETING
- CREATING → FAILED

Limitations on using partition indexes

Once you have created a partition index, note these changes to table and partition functionality:

New partition creation (after Index Addition)

After a partition index is created on a table, all new partitions added to the table will be validated for the data type checks for indexed keys. The partition value of the indexed keys will be validated for data type format. If the data type check fails, the create partition operation will fail. For the *sales_data* table, if an index is created for keys (category, year) where the category is of type `string` and year of type `int`, the creation of the new partition with a value of YEAR as "foo" will fail.

After indexes are enabled, the addition of partitions with indexed key values having the characters U+0000, U+00001, and U+0002 will start to fail.

Table updates

Once a partition index is created on a table, you cannot modify the partition key names for existing partition keys, and you cannot change the type, or order, of keys which are registered with the index.

Using indexes for an optimized GetPartitions call

When you call `GetPartitions` on a table with an index, you can include an expression, and if applicable the Data Catalog will use an index if possible. The first key of the index should be passed in the expression for the indexes to be used in filtering. Index optimization in filtering is applied as a best effort. The Data Catalog tries to use index optimization as much as possible, but in case of a missing index, or unsupported operator, it falls back to the existing implementation of loading all partitions.

For the *sales_data* table above, lets add the index [Country, Category, Year]. If "Country" is not passed in the expression, the registered index will not be able to filter partitions using indexes. You can add up to 3 indexes to support various query patterns.

Lets take some example expressions and see how indexes work on them:

Expressions	How index will be used
Country = 'US'	Index will be used to filter partitions.
Country = 'US' and Category = 'Shoes'	Index will be used to filter partitions.
Category = 'Shoes'	Indexes will not be used as "country" is not provided in the expression. All partitions will be loaded to return a response.
Country = 'US' and Category = 'Shoes' and Year > '2018'	Index will be used to filter partitions.
Country = 'US' and Category = 'Shoes' and Year > '2018' and month = 2	Index will be used to fetch all partitions with country = "US" and category = "shoes" and year > 2018. Then, filtering on the month expression will be performed.
Country = 'US' AND Category = 'Shoes' OR Year > '2018'	Indexes will not be used as an `OR` operator is present in the expression.
Country = 'US' AND Category = 'Shoes' AND (Year = 2017 OR Year = '2018')	Index will be used to fetch all partitions with country = "US" and category = "shoes", and then filtering on the year expression will be performed.
Country in ('US', 'UK') AND Category = 'Shoes'	Indexes will not be used for filtering as the `IN` operator is not supported currently.
Country = 'US' AND Category in ('Shoes', 'Books')	Index will be used to fetch all partitions with country = "US", and then filtering on the Category expression will be performed.

Expressions	How index will be used
Country = 'US' AND Category in ('Shoes', 'Books') AND (creationDate > '2023-9-01'	Index will be used to fetch all partition s with country = "US", with creationDate > '2023-9-01', and then filtering on the Category expression will be performed.

Integration with engines

Redshift Spectrum, Amazon EMR and AWS Glue ETL Spark DataFrames are able to utilize indexes for fetching partitions after indexes are in an ACTIVE state in AWS Glue. Athena and AWS Glue ETL Dynamic frames require you to follow extra steps to utilize indexes for query improvement.

Enable partition filtering

To enable partition filtering in Athena, you need to update the table properties as follows:

1. In the AWS Glue console, under **Data Catalog**, choose **Tables**.
2. Choose a table.
3. Under **Actions**, choose **Edit table**.
4. Under **Table properties**, add the following:
   - Key –`partition_filtering.enabled`
   - Value – `true`
5. Choose **Apply**.

Alternatively, you can set this parameter by running an ALTER TABLE SET PROPERTIES query in Athena.

```
ALTER TABLE partition_index.table_with_index
SET TBLPROPERTIES ('partition_filtering.enabled' = 'true')
```

Updating manually created Data Catalog tables using crawlers

You might want to create AWS Glue Data Catalog tables manually and then keep them updated with AWS Glue crawlers. Crawlers running on a schedule can add new partitions and update the tables with any schema changes. This also applies to tables migrated from an Apache Hive metastore.

To do this, when you define a crawler, instead of specifying one or more data stores as the source of a crawl, you specify one or more existing Data Catalog tables. The crawler then crawls the data stores specified by the catalog tables. In this case, no new tables are created; instead, your manually created tables are updated.

The following are other reasons why you might want to manually create catalog tables and specify catalog tables as the crawler source:

- You want to choose the catalog table name and not rely on the catalog table naming algorithm.
- You want to prevent new tables from being created in the case where files with a format that could disrupt partition detection are mistakenly saved in the data source path.

For more information, see Step 2: Choose data sources and classifiers.

Data Catalog table properties

Table properties, or parameters, as they are known in the AWS CLI, are unvalidated key and value strings. You can set your own properties on the table to support uses of the Data Catalog outside of AWS Glue. Other services using the Data Catalog may do so as well. AWS Glue sets some table properties when running jobs or crawlers. Unless otherwise described, these properties are for internal use, we do not support that they will continue to exist in their current form, or support product behavior if these properties are manually changed.

For more information about table properties set by AWS Glue crawlers, see the section called “Parameters set on Data Catalog tables by crawler”.

Integrating with other AWS services

While you can use AWS Glue crawlers to populate the AWS Glue Data Catalog, there are several AWS services that can automatically integrate with and populate the catalog for you. The following sections provide more information about the specific use cases supported by AWS services that can populate the Data Catalog.

Topics

- AWS Lake Formation

- Amazon Athena

AWS Lake Formation

AWS Lake Formation is a service that makes it easier to set up a secure data lake in AWS. Lake Formation is built on AWS Glue, and Lake Formation and AWS Glue share the same AWS Glue Data Catalog. You can register your Amazon S3 data location with Lake Formation, and use Lake Formation console to create databases and tables in the AWS Glue Data Catalog, define data access policies, and audit data access across your data lake from a central place. You can use the Lake Formation fine-grained access control to manage your existing Data Catalog resources and Amazon S3 data locations.

With data registered with Lake Formation, you can securely share Data Catalog resources across IAM principals, AWS accounts, AWS organizations, and organizational units.

For more information about creating Data Catalog resources using Lake Formation, see Creating Data Catalog tables and databases in the AWS Lake Formation Developer Guide.

Amazon Athena

Amazon Athena uses the Data Catalog to store and retrieve table metadata for the Amazon S3 data in your AWS account. The table metadata lets the Athena query engine know how to find, read, and process the data that you want to query.

You can populate the AWS Glue Data Catalog by using Athena `CREATE TABLE` statements directly. You can manually define and populate the schema and partition metadata in the Data Catalog without needing to run a crawler.

1. In the Athena console, create a database that will store the table metadata in the Data Catalog.
2. Use the `CREATE EXTERNAL TABLE` statement to define the schema of your data source.
3. Use the `PARTITIONED BY` clause to define any partition keys if your data is partitioned.
4. Use the `LOCATION` clause to specify the Amazon S3 path where your actual data files are stored.
5. Run the `CREATE TABLE` statement.

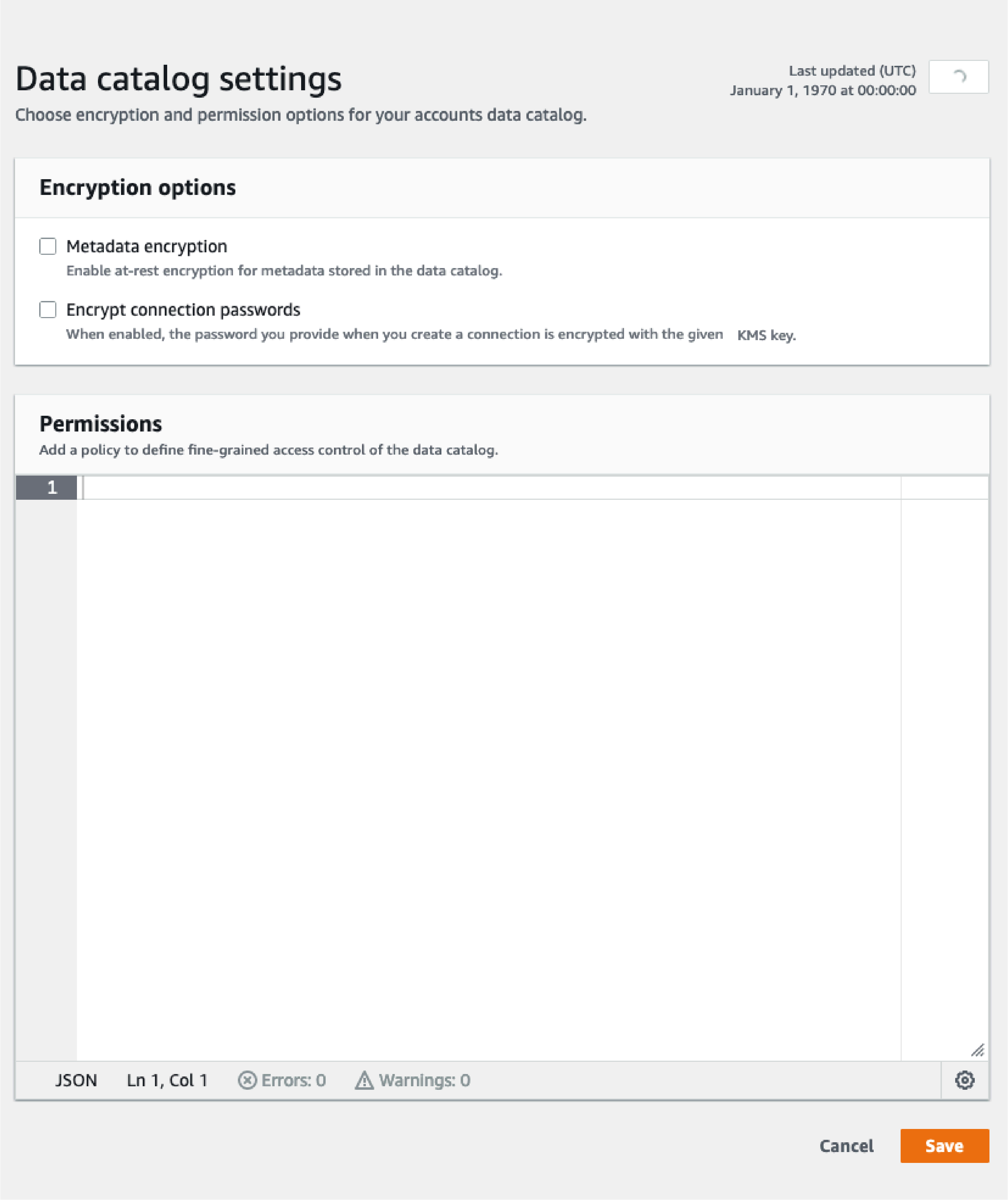
   This query creates the table metadata in the Data Catalog based on your defined schema and partitions, without actually crawling the data.

You can query the table in Athena, and it will use the metadata from the Data Catalog to access and query your data files in Amazon S3.

For more information, see Creating databases and tables in the Amazon Athena User Guide.

Data Catalog settings

The Data Catalog settings contains options to set encryption and permissions options for the Data Catalog in your account.



Data catalog settings

Choose encryption and permission options for your accounts data catalog.

Last updated (UTC)
January 1, 1970 at 00:00:00

Encryption options

☐ Metadata encryption
Enable at-rest encryption for metadata stored in the data catalog.

☐ Encrypt connection passwords
When enabled, the password you provide when you create a connection is encrypted with the given KMS key.

Permissions

Add a policy to define fine-grained access control of the data catalog.

To change the fine-grained access control of the Data Catalog

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. Choose an encryption option.

    - **Metadata encryption** – Select this check box to encrypt the metadata in your Data Catalog. Metadata is encrypted at rest using the AWS Key Management Service (AWS KMS) key that you specify. For more information, see Encrypting your Data Catalog.
    - **Encrypt connection passwords** – Select this check box to encrypt passwords in the AWS Glue connection object when the connection is created or updated. Passwords are encrypted using the AWS KMS key that you specify. When passwords are returned, they are encrypted. This option is a global setting for all AWS Glue connections in the Data Catalog. If you clear this check box, previously encrypted passwords remain encrypted using the key that was used when they were created or updated. For more information about AWS Glue connections, see Connecting to data.

      When you enable this option, choose an AWS KMS key, or choose **Enter a key ARN** and provide the Amazon Resource Name (ARN) for the key. Enter the ARN in the form `arn:aws:kms:region:account-id:key/key-id`. You can also provide the ARN as a key alias, such as `arn:aws:kms:region:account-id:alias/alias-name`.

      > **Important**
      >
      > If this option is selected, any user or role that creates or updates a connection must have `kms:Encrypt` permission on the specified KMS key.

      For more information, see Encrypting connection passwords.
3. Choose **Settings**, and then in the **Permissions** editor, add the policy statement to change fine-grained access control of the Data Catalog for your account. Only one policy at a time can be attached to a Data Catalog. You can paste a JSON resource policy into this control. For more information, see Resource-based policies within AWS Glue.
4. Choose **Save** to update your Data Catalog with any changes you made.

You can also use AWS Glue API operations to put, get, and delete resource policies. For more information, see Security APIs in AWS Glue.

Populating and managing transactional tables

Apache Iceberg, Apache Hudi, and Linux Foundation Delta Lake are open-source table formats designed for handling large-scale data analytics and data lake workloads in Apache Spark.

You can populate Iceberg, Hudi, and Delta Lake tables in the AWS Glue Data Catalog using the following methods:

- AWS Glue crawler; – AWS Glue crawlers can automatically discover and populate Iceberg, Hudi and Delta Lake table metadata in the Data Catalog. For more information, see Using crawlers to populate the Data Catalog .
- AWS Glue ETL Jobs – You can create ETL jobs to write data to Iceberg, Hudi, and Delta Lake tables and populate their metadata in the Data Catalog. For more information, see Using data lake frameworks with AWS Glue ETL jobs.
- AWS Glue console, AWS Lake Formation console, AWS CLI or API – You can use the AWS Glue console, Lake Formation console, or API to create and manage Iceberg table definitions in the Data Catalog.

Topics

- Creating Apache Iceberg tables
- Optimizing Iceberg tables
- Optimizing query performance for Iceberg tables

Creating Apache Iceberg tables

You can create Apache Iceberg tables that use the Apache Parquet data format in the AWS Glue Data Catalog with data residing in Amazon S3. A table in the Data Catalog is the metadata definition that represents the data in a data store. By default, AWS Glue creates Iceberg v2 tables. For the difference between v1 and v2 tables, see Format version changes in the Apache Iceberg documentation.

Apache Iceberg is an open table format for very large analytic datasets. Iceberg allows for easy changes to your schema, also known as schema evolution, meaning that users can add, rename, or remove columns from a data table without disrupting the underlying data. Iceberg also provides support for data versioning, which allows users to track changes to data overtime. This enables the time travel feature, which allows users to access and query historical versions of data and analyze changes to the data between updates and deletes.

You can use AWS Glue or Lake Formation console or the `CreateTable` operation in the AWS Glue API to create an Iceberg table in the Data Catalog. For more information, see CreateTable action (Python: create_table).

When you create an Iceberg table in the Data Catalog, you must specify the table format and metadata file path in Amazon S3 to be able to perform reads and writes.

You can use Lake Formation to secure your Iceberg table using fine-grained access control permissions when you register the Amazon S3 data location with AWS Lake Formation. For source data in Amazon S3 and metadata that is not registered with Lake Formation, access is determined by IAM permissions policies for Amazon S3 and AWS Glue actions. For more information, see Managing permissions.

> **Note**
>
> Data Catalog doesn’t support creating partitions and adding Iceberg table properties.

Prerequisites

To create Iceberg tables in the Data Catalog, and set up Lake Formation data access permissions, you need to complete the following requirements:

1. **Permissions required to create Iceberg tables without the data registered with Lake Formation.**

   In addition to the permissions required to create a table in the Data Catalog, the table creator requires the following permissions:

   - `s3:PutObject` on resource arn:aws:s3:::{bucketName}
   - `s3:GetObject` on resource arn:aws:s3:::{bucketName}
   - `s3:DeleteObject`on resource arn:aws:s3:::{bucketName}
2. **Permissions required to create Iceberg tables with data registered with Lake Formation:**

   To use Lake Formation to manage and secure the data in your data lake, register your Amazon S3 location that has the data for tables with Lake Formation. This is so that Lake Formation can vend credentials to AWS analytical services such as Athena, Redshift Spectrum, and Amazon EMR to access data. For more information on registering an Amazon S3 location, see Adding an Amazon S3 location to your data lake.

A principal who reads and writes the underlying data that is registered with Lake Formation requires the following permissions:

- `lakeformation:GetDataAccess`
- `DATA_LOCATION_ACCESS`

  A principal who has data location permissions on a location also has location permissions on all child locations.

  For more information on data location permissions, see Underlying data access controlulink.

To enable compaction, the service needs to assume an IAM role that has permissions to update tables in the Data Catalog. For details, see Table optimization prerequisites

Creating an Iceberg table

You can create Iceberg v1 and v2 tables using AWS Glue or Lake Formation console or AWS Command Line Interface as documented on this page. You can also create Iceberg tables using the AWS Glue crawler. For more information, see Data Catalog and Crawlers in the AWS Glue Developer Guide.

To create an Iceberg table

Console

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. Under Data Catalog, choose **Tables**, and use the **Create table** button to specify the following attributes:
   - **Table name –** Enter a name for the table. If you're using Athena to access tables, use these naming tips in the Amazon Athena User Guide.
   - **Database –** Choose an existing database or create a new one.
   - **Description –** The description of the table. You can write a description to help you understand the contents of the table.

- **Table format** – For **Table format**, choose Apache Iceberg.
- **Enable compaction** – Choose **Enable compaction** to compact small Amazon S3 objects in the table into larger objects.
- **IAM role** – To run compaction, the service assumes an IAM role on your behalf. You can choose an IAM role using the drop-down. Ensure that the role has the permissions required to enable compaction.

  To learn more about the required permissions, see Table optimization prerequisites .
- **Location** – Specify the path to the folder in Amazon S3 that stores the metadata table. Iceberg needs a metadata file and location in the Data Catalog to be able to perform reads and writes.
- **Schema** – Choose **Add columns** to add columns and data types of the columns. You have the option to create an empty table and update the schema later. Data Catalog supports Hive data types. For more information, see Hive data types.

  Iceberg allows you to evolve schema and partition after you create the table. You can use Athena queries to update the table schema and Spark queries for updating partitions.

AWS CLI

```
aws glue create-table \
    --database-name iceberg-db \
    --region us-west-2 \
    --open-table-format-input '{
      "IcebergInput": {
           "MetadataOperation": "CREATE",
           "Version": "2"
         }
      }' \
    --table-input '{"Name":"test-iceberg-input-demo",
            "TableType": "EXTERNAL_TABLE",
            "StorageDescriptor":{
               "Columns":[
                   {"Name":"col1", "Type":"int"},
                   {"Name":"col2", "Type":"int"},
                   {"Name":"col3", "Type":"string"}
                ],
               "Location":"s3://DOC_EXAMPLE_BUCKET_ICEBERG/"
            }
```

```
        }'
```

Optimizing Iceberg tables

AWS Glue supports mutiple table optimization options to enhance the management and performance of Apache Iceberg tables used by the AWS analytical engines and ETL jobs. These optimizers provide efficient storage utilization, improved query performance, and effective data management. There are three types of table optimizers available in AWS Glue:

- **Compaction –** Data compaction compacts small data files to reduce storage usage and improve read performance. Data files are merged and rewritten to remove obsolete data and consolidate fragmented data into larger, more efficient files. You can configure compaction to run automatically.

  Binpack is the default compaction strategy in Apache Iceberg. It combines smaller data files into larger ones for optimal performance. Compaction also supports sort and Z-order strategies that cluster similar data together. Sort organizes data based on specified columns, improving query performance for filtered operations. Z-order creates sorted datasets that enhance query performance when multiple columns are queried simultaneously. All three compaction strategies - bincpak, sort, and Z-order - reduce the amount of data scanned by query engines, thereby lowering query processing costs.
- **Snapshot retention –** Snapshots are timestamped versions of an Iceberg table. Snapshot retention configurations allow customers to enforce how long to retain snapshots and how many snapshots to retain. Configuring a snapshot retention optimizer can help manage storage overhead by removing older, unnecessary snapshots and their associated underlying files.
- **Orphan file deletion –** Orphan files are files that are no longer referenced by the Iceberg table metadata. These files can accumulate over time, especially after operations like table deletions or failed ETL jobs. Enabling orphan file deletion allows AWS Glue to periodically identify and remove these unnecessary files, freeing up storage.

Catalog-level optimization configuration is available through the Lake Formation console and using the AWS Glue `UpdateCatalog` API operation. You can enable or disable compaction, snapshot retention, and orphan file deletion optimizers for individual Iceberg tables in the Data Catalog using the AWS Glue console, AWS CLI, or AWS Glue API operations.

Topics

- Table optimization prerequisites
- Catalog-level table optimizers
- Compaction optimization
- Snapshot retention optimization
- Deleting orphan files
- Viewing optimization details
- Viewing Amazon CloudWatch metrics
- Deleting an optimizer
- Considerations and limitations
- Supported Regions for table optimizers

Table optimization prerequisites

The table optimizer assumes the permissions of the AWS Identity and Access Management (IAM) role that you specify when you enable optimization options (compaction, snapshot retention, and orphan file delettion) for a table. You can either create s single role for all optimizers or create separate roles for each optimizer.

> **Note**
>
> The orphan file deletion optimizer doesn't require the `glue:updateTable` or `s3:putObject` permissions. The snapshot expiration and compaction optimizers require the same set of permissions.

The IAM role must have the permissions to read data and update metadata in the Data Catalog. You can create an IAM role and attach the following inline policies:

- Add the following inline policy that grants Amazon S3 read/write permissions on the location for data that is not registered with AWS Lake Formation. This policy also includes permissions to update the table in the Data Catalog, and to permit AWS Glue to add logs in Amazon CloudWatch logs and publish metrics. For source data in Amazon S3 that isn't registered with Lake Formation, access is determined by IAM permissions policies for Amazon S3 and AWS Glue actions.

In the following inline policies, replace `bucket-name` with your Amazon S3 bucket name, `aws-account-id` and `region` with a valid AWS account number and Region of the Data Catalog, `database_name` with the name of your database, and `table_name` with the name of the table.

JSON

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "s3:PutObject",
                "s3:GetObject",
                "s3:DeleteObject"
            ],
            "Resource": [
                "arn:aws:s3:::amzn-s3-demo-bucket/*"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "s3:ListBucket"
            ],
            "Resource": [
                "arn:aws:s3:::amzn-s3-demo-bucket"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "glue:UpdateTable",
                "glue:GetTable"
            ],
            "Resource": [
                "arn:aws:glue:us-east-1:111122223333:table/<database-
name>/<table-name>",
                "arn:aws:glue:us-east-1:111122223333:database/<database-name>",
                "arn:aws:glue:us-east-1:111122223333:catalog"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "logs:CreateLogGroup",
                "logs:CreateLogStream",
                "logs:PutLogEvents"
            ],
            "Resource": [
                "arn:aws:logs:us-east-1:111122223333:log-group:/aws-glue/
iceberg-compaction/logs:*",
                "arn:aws:logs:us-east-1:111122223333:log-group:/aws-glue/
iceberg-retention/logs:*",
                "arn:aws:logs:us-east-1:111122223333:log-group:/aws-glue/
iceberg-orphan-file-deletion/logs:*"
            ]
        }
    ]
}
```

- Use the following policy to enable compaction for data registered with Lake Formation.

  If the optimization role doesn't have `IAM_ALLOWED_PRINCIPALS` group permissions granted on the table, the role requires Lake Formation `ALTER`, `DESCRIBE`, `INSERT` and `DELETE` permissions on the table.

  For more information on registering an Amazon S3 bucket with Lake Formation, see Adding an Amazon S3 location to your data lake.

  JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "lakeformation:GetDataAccess"
      ],
      "Resource": "*"
    },
    {
      "Effect": "Allow",
      "Action": [
        "glue:UpdateTable",
        "glue:GetTable"
      ],
      "Resource": [
        "arn:aws:glue:us-east-1:111122223333:table/databaseName/tableName",
        "arn:aws:glue:us-east-1:111122223333:database/databaseName",
        "arn:aws:glue:us-east-1:111122223333:catalog"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "logs:CreateLogGroup",
        "logs:CreateLogStream",
        "logs:PutLogEvents"
      ],
      "Resource": [
        "arn:aws:logs:us-east-1:111122223333:log-group:/aws-glue/iceberg-
compaction/logs:*",
        "arn:aws:logs:us-east-1:111122223333:log-group:/aws-glue/iceberg-
retention/logs:*",
        "arn:aws:logs:us-east-1:111122223333:log-group:/aws-glue/iceberg-
orphan-file-deletion/logs:*"
      ]
    }
  ]
}
```

- (Optional) To optimize Iceberg tables with data in Amazon S3 buckets encrypted using Server-side encryption, the compaction role requires permissions to decrypt Amazon S3 objects and generate a new data key to write objects to the encrypted buckets. Add the following policy to the desired AWS KMS key. We support only bucket-level encryption.

```
{
    "Effect": "Allow",
    "Principal": {
        "AWS": "arn:aws:iam::<aws-account-id>:role/<optimizer-role-name>"
    },
    "Action": [
        "kms:Decrypt",
        "kms:GenerateDataKey"
    ],
    "Resource": "*"
}
```

- (Optional) For data location registered with Lake Formation, the role used to register the location requires permissions to decrypt Amazon S3 objects and generate a new data key to write objects to the encrypted buckets. For more information, see Registering an encrypted Amazon S3 location.
- (Optional) If the AWS KMS key is stored in a different AWS account, you need to include the following permissions to the compaction role.

  JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "kms:Decrypt",
        "kms:GenerateDataKey"
      ],
      "Resource": [
        "arn:aws:kms:us-east-1:111122223333:key/key-id"
      ]
    }
  ]
}
```

- The role you use to run compaction must have the `iam:PassRole` permission on the role.

  JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "iam:PassRole"
      ],
      "Resource": [
        "arn:aws:iam::111122223333:role/<optimizer-role-name>"
      ]
    }
  ]
}
```

- Add the following trust policy to the role for AWS Glue service to assume the IAM role to run the compaction process.

  JSON

  ```
  {
    "Version": "2012-10-17",
    "Statement": [
      {
        "Sid": "",
        "Effect": "Allow",
        "Principal": {
          "Service": "glue.amazonaws.com"
        },
        "Action": "sts:AssumeRole"
      }
    ]
  }
  ```

- (Optional) To update the Data Catalog settings to enable catalog-level table optimizations, the IAM role used must have the `glue:UpdateCatalog` permission or AWS Lake Formation `ALTER CATALOG` permission on the root catalog. You can use `GetCatalog` API to verify the catalog properties.

Catalog-level table optimizers

With a one-time catalog configuration, you can set up automatic optimizers such as compaction, snapshot retention, and orphan file deletion for all new and updated Apache Iceberg tables in the AWS Glue Data Catalog. Catalog-level optimizer configurations allow you to apply consistent optimizer settings across all tables within a catalog, eliminating the need to configure optimizers individually for each table.

Data lake administrators can configure the table optimizers by selecting the default catalog in the Lake Formation console and enabling optimizers using the `Table optimization` option. When you create new tables or update existing tables in the Data Catalog, the Data Catalog automatically runs the table optimizations to reduce operational burden.

If you have configured optimization at the table level or if you have previously deleted the table optimization settings for a table, those table-specific settings take precedence over the default catalog settings for table optimization. If a configuration parameter is not defined at either the table or catalog level, the Iceberg table property value will be applied. This setting is applicable to snapshot retention and orphan file deletion optimizer.

When enabling catalog-level optimizers, consider the following:

- When you configure optimization settings at the time of catalog creation and subsequently disable the optimizations through an Update Catalog request, the operation will cascade through all the tables within the catalog.
- If you have already configured optimizers for a given table, then the disable operation at the catalog level will not impact this table.
- When you disable optimizers at the catalog level, tables with existing optimizer configurations will maintain their specific settings and remain unaffected by the catalog-level change. However, tables without their own optimizer configurations will inherit the disabled state from the catalog level.
- Since snapshot retention and orphan file deletion optimizers can be schedule-based, updates will introduce a random delay to the start of their schedule. This will cause each optimizer to start at slightly different times, spreading out the load and reducing the likelihood of exceeding service limits.

Topics

- Enabling catalog-level automatic table optimization
- Viewing catalog-level optimizations
- Disabling catalog-level table optimization

Enabling catalog-level automatic table optimization

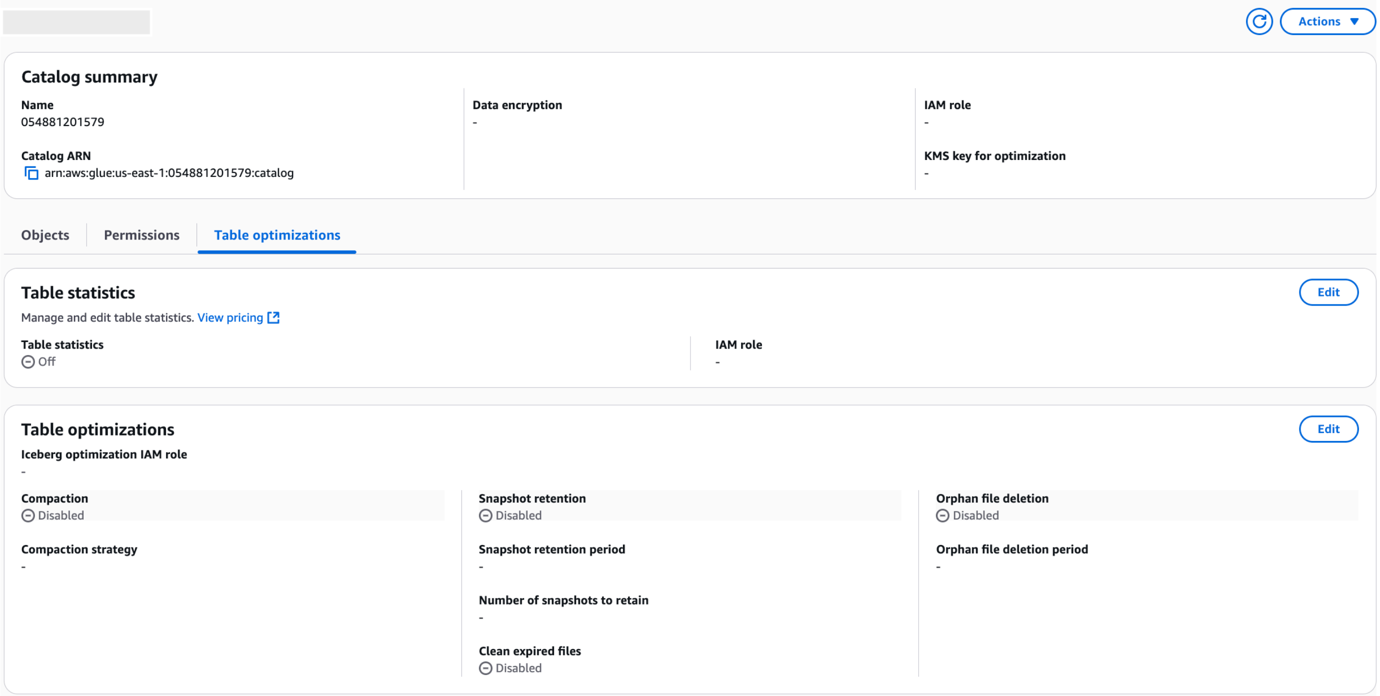
You can enable the automatic table optimization for all new Apache Iceberg tables in the Data Catalog. After creating the table, you can also explicitly update the table optimization settings manually.

To update the Data Catalog settings to enable catalog-level table optimizations, the IAM role used must have the `glue:UpdateCatalog` permission on the root catalog. You can use `GetCatalog` API to verify the catalog properties.

For the Lake Formation managed tables, the IAM role selected during the catalog optimization configuration requires Lake Formation `ALTER`, `DESCRIBE`, `INSERT`, and `DELETE` permissions for any new tables or updated tables.

To enable catalog-level optimizers (console)

1. Open the Lake Formation console at https://console.aws.amazon.com/lakeformation/.
2. In the navigation pane, choose **Data Catalog**.
3. Select the **Catalogs** tab.
4. Choose the account-level catalog.
5. Choose **Table optimizations**, **Edit** under **Table optimizations** tab. You can also choose **Edit optimizations** from **Actions**.



Catalog summary

Name
054881201579

Catalog ARN
arn:aws:glue:us-east-1:054881201579:catalog

Data encryption
-

IAM role
-

KMS key for optimization
-

Table statistics

Manage and edit table statistics. View pricing

Table statistics
Off

IAM role
-

Table optimizations

Iceberg optimization IAM role
-

Compaction
Disabled

Compaction strategy
-

Snapshot retention
Disabled

Snapshot retention period
-

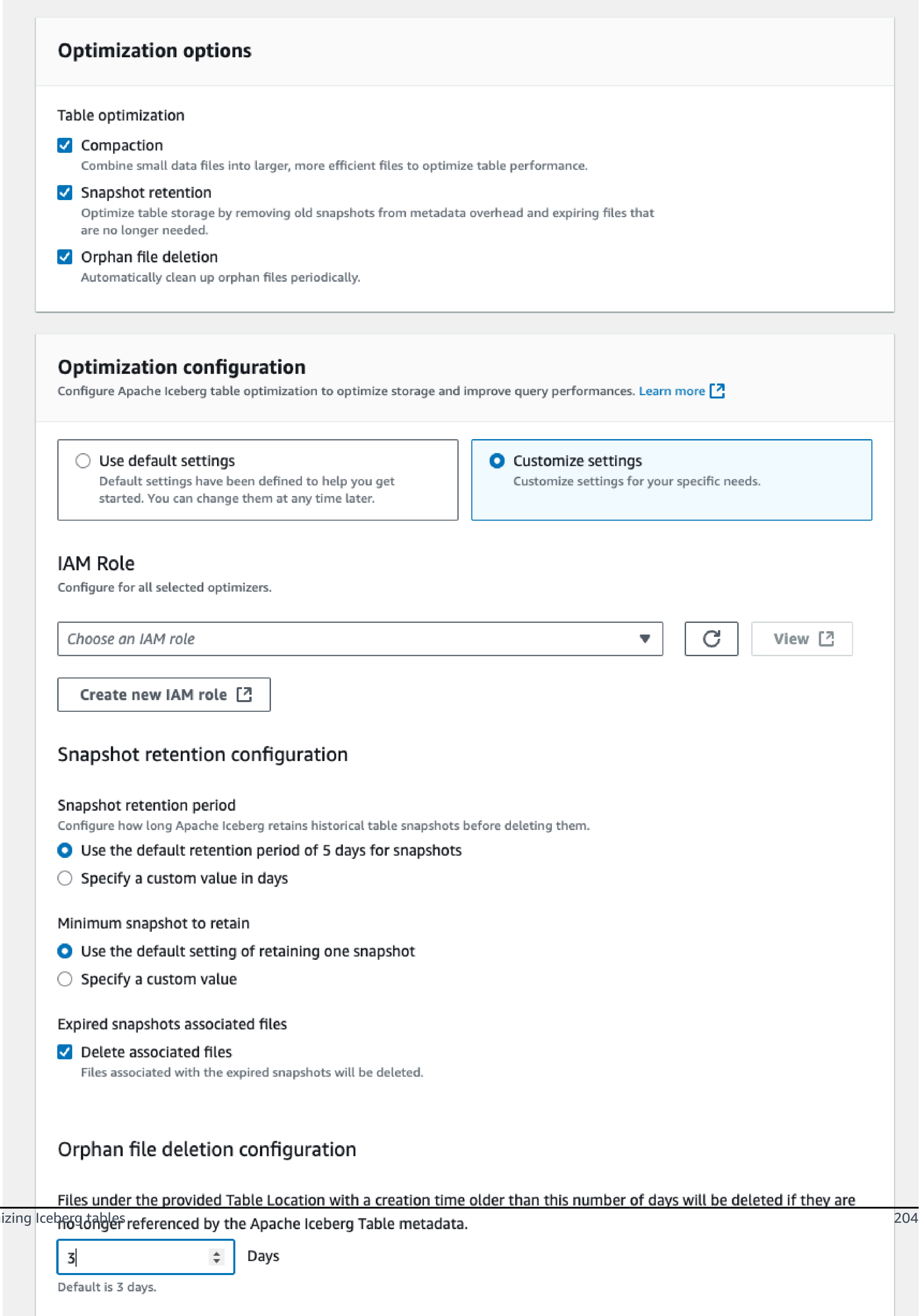
Number of snapshots to retain
-

Clean expired files
Disabled

Orphan file deletion
Disabled

Orphan file deletion period
-

6. On the **Table optimization** page, configure the following options:



Optimization options

Table optimization

☑ Compaction
Combine small data files into larger, more efficient files to optimize table performance.

☑ Snapshot retention
Optimize table storage by removing old snapshots from metadata overhead and expiring files that are no longer needed.

☑ Orphan file deletion
Automatically clean up orphan files periodically.

Optimization configuration

Configure Apache Iceberg table optimization to optimize storage and improve query performances. Learn more

○ Use default settings
Default settings have been defined to help you get started. You can change them at any time later.

● Customize settings
Customize settings for your specific needs.

IAM Role

Configure for all selected optimizers.

Snapshot retention configuration

Snapshot retention period
Configure how long Apache Iceberg retains historical table snapshots before deleting them.

● Use the default retention period of 5 days for snapshots
○ Specify a custom value in days

Minimum snapshot to retain

● Use the default setting of retaining one snapshot
○ Specify a custom value

Expired snapshots associated files

☑ Delete associated files
Files associated with the expired snapshots will be deleted.

Orphan file deletion configuration

Files under the provided Table Location with a creation time older than this number of days will be deleted if they are no longer referenced by the Apache Iceberg Table metadata.

3 Days

Default is 3 days.

   a. Configure **Compaction** settings:

      - Enable/disable compaction.
      - Choose the IAM role that has the necessary permissions to run the optimizers.

        For more information on the permission requirements for the IAM role, see [Table optimization prerequisites](#).

   b. Configure **Snapshot retention** settings:

      - Enable/disable retention.
      - Set snapshot retention period in days - default is 5 days.
      - Set number of snapshots to retain - default is 1 snapshot.
      - Enable/disable cleaning of expired files.

   c. Configure **Orphan file deletion** settings:

      - Enable/disable orphan file deletion.
      - Set orphan file retention period in days - default is 3 days.

7. Choose **Save**.

Enabling Catalog-Level Optimizers via AWS CLI

Use the following CLI command to update an existing catalog with optimizer settings:

Example Update catalog with optimizer settings

```
aws glue update-catalog \
   --name catalog-id \
  --catalog-input \
  '{
    "CatalogId": "111122223333",
    "CatalogInput": {
        "CatalogProperties": {
            "CustomProperties": {
                "ColumnStatistics.Enabled": "false",
                "ColumnStatistics.RoleArn": "arn:aws:iam::111122223333:role/service-
role/stats-role-name"
            },
            "IcebergOptimizationProperties": {
                "RoleArn": "arn:aws:iam::111122223333:role/optimizer-role-name",
                 "Compaction": {
                     "enabled": "true"
                 },
                 "Retention": {
                     "enabled": "true",
                     "snapshotRetentionPeriodInDays": "10",
                     "numberOfSnapshotsToRetain": "5",
                     "cleanExpiredFiles": "true"
                 },
                 "OrphanFileDeletion": {
                     "enabled": "true",
                     "orphanFileRetentionPeriodInDays": "3"
                 }
             }
         }
     }
}'
```

If you encounter issues with catalog-level optimizers, check the following:

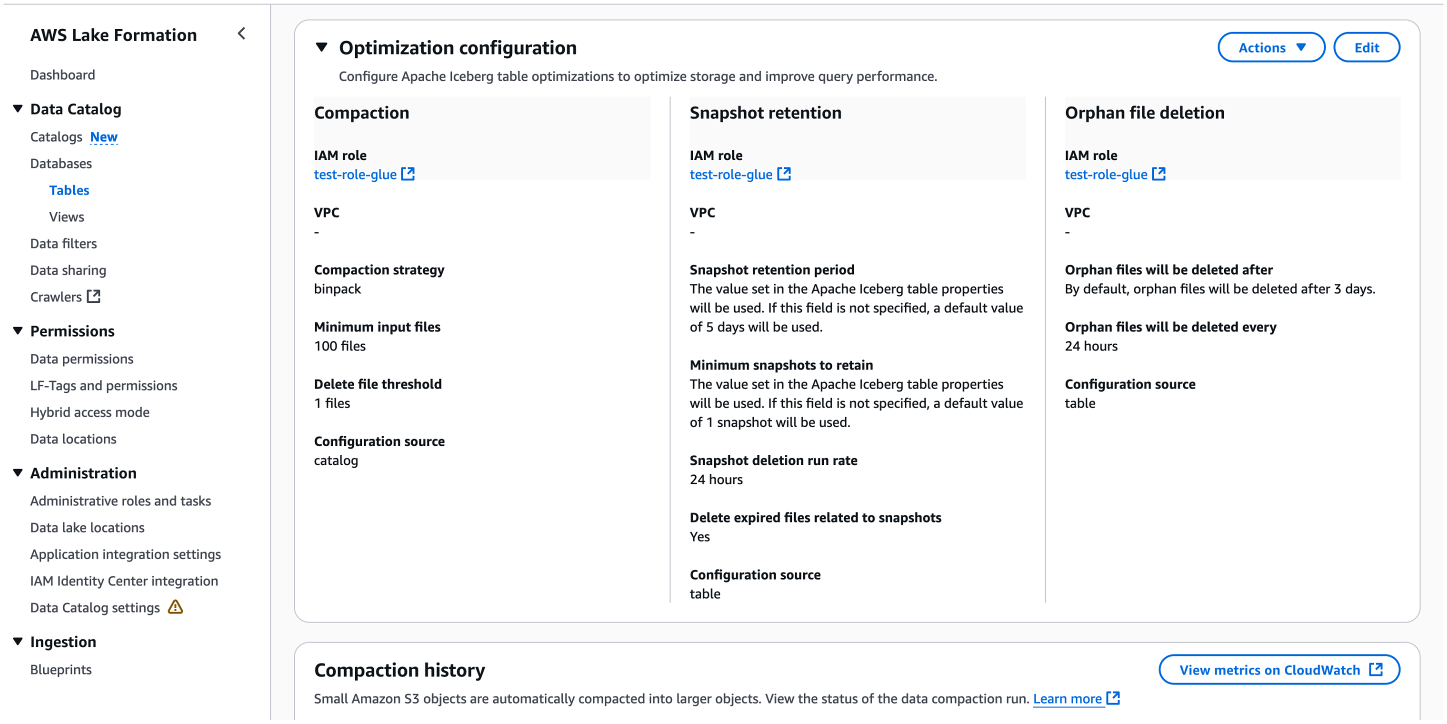
- Ensure the IAM role has the correct permissions as outlined in the Prerequisites section.
- Check CloudWatch logs for any error messages related to optimizer operations.

  For more information, see View available metrics in the *Amazon CloudWatch User Guide*.
- Verify that the catalog settings were successfully applied by checking the catalog configuration.
- For table access failures, check the CloudWatch logs and EventBridge notifications for detailed error information.

Viewing catalog-level optimizations

When catalog-level table optimization is enabled, anytime an Apache Iceberg table is created or updated via the `CreateTable` or `UpdateTable` APIs through AWS Management Console, SDK, or AWS Glue crawler, an equivalent table level setting is created for that table.

After you create or update a table, you can verify the table details to confirm the table optimization. The `Table optimization` shows the `Configuration source` property set as `Catalog`.


AWS Lake Formation
Dashboard
Data Catalog
Catalogs New
Databases
Tables
Views
Data filters
Data sharing
Crawlers
Permissions
Data permissions
LF-Tags and permissions
Hybrid access mode
Data locations
Administration
Administrative roles and tasks
Data lake locations
Application integration settings
IAM Identity Center integration
Data Catalog settings
Ingestion
Blueprints
Actions
Edit
View metrics on CloudWatch
Learn more

Optimization configuration

Configure Apache Iceberg table optimizations to optimize storage and improve query performance.

Compaction

IAM role
test-role-glue

VPC
-

Compaction strategy
binpack

Minimum input files
100 files

Delete file threshold
1 files

Configuration source
catalog

Snapshot retention

IAM role
test-role-glue

VPC
-

Snapshot retention period
The value set in the Apache Iceberg table properties will be used. If this field is not specified, a default value of 5 days will be used.

Minimum snapshots to retain
The value set in the Apache Iceberg table properties will be used. If this field is not specified, a default value of 1 snapshot will be used.

Snapshot deletion run rate
24 hours

Delete expired files related to snapshots
Yes

Configuration source
table

Orphan file deletion

IAM role
test-role-glue

VPC
-

Orphan files will be deleted after
By default, orphan files will be deleted after 3 days.

Orphan files will be deleted every
24 hours

Configuration source
table

Compaction history

Small Amazon S3 objects are automatically compacted into larger objects. View the status of the data compaction run. Learn more

Disabling catalog-level table optimization

You can disable table optimization for new tables using the AWS Lake Formation console, the `glue:UpdateCatalog` API.

To disable the table optimizations at the catalog level

1. Open the Lake Formation console at https://console.aws.amazon.com/lakeformation/.
2. On the left navigation bar, choose **Catalogs**.
3. On the **Catalog summary** page, choose **Edit** under **Table optimizations**.
4. On the **Edit optimization** page, unselect the **Optimization options**.
5. Choose **Save**.

Compaction optimization

The Amazon S3 data lakes using open table formats like Apache Iceberg store data as S3 objects. Having thousands of small Amazon S3 objects in a data lake table increases metadata overhead and affects read performance. AWS Glue Data Catalog provides managed compaction for Iceberg tables, compacting small objects into larger ones for better read performance by AWS analytics services like Amazon Athena and Amazon EMR, and AWS Glue ETL jobs. Data Catalog performs compaction without interfering with concurrent queries and supports compaction only for Parquet format tables.

The table optimizer continuously monitors table partitions and kicks off the compaction process when the threshold is exceeded for the number of files and file sizes.

In the Data Catalog, the compaction process starts when a table or any of its partitions have more than 100 files. Each file must be smaller than 75% of the target file size. The target file size is defined by the `write.target-file-size-bytes` table property, which defaults to 512 MB if not explicitly set.

For limitations, see Supported formats and limitations for managed data compaction .

Topics

- Enabling compaction optimizer
- Disabling compaction optimizer

Enabling compaction optimizer

You can use AWS Glue console, AWS CLI, or AWS API to enable compaction for your Apache Iceberg tables in the AWS Glue Data Catalog. For new tables, you can choose Apache Iceberg as table format and enable compaction when you create the table. Compaction is disabled by default for new tables.

Console

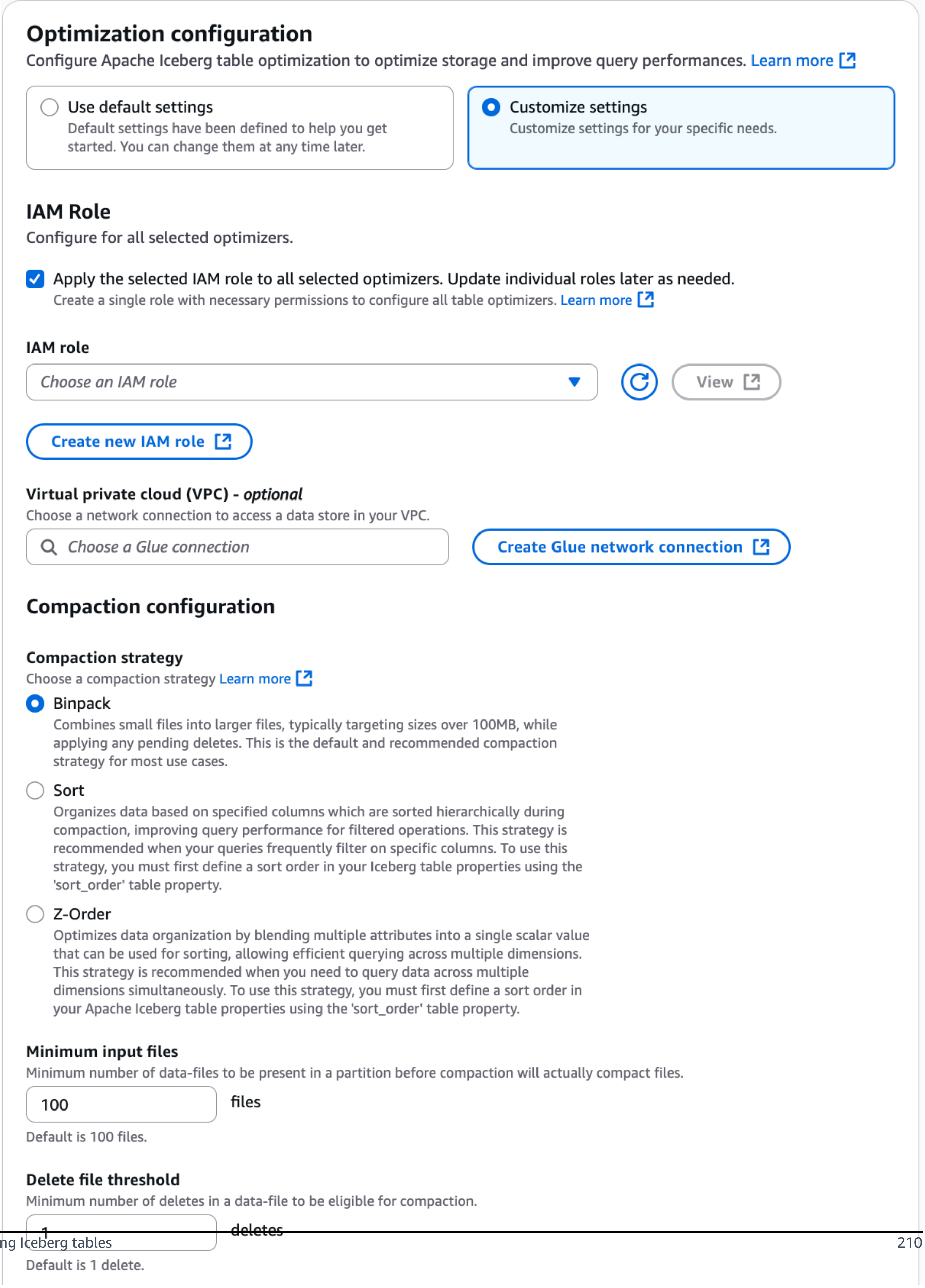
To enable compaction

1. Open the AWS Glue console at https://console.aws.amazon.com/glue/ and sign in as a data lake administrator, the table creator, or a user who has been granted the `glue:UpdateTable` and `lakeformation:GetDataAccess` permissions on the table.
2. In the navigation pane, under **Data Catalog**, choose **Tables**.
3. On the **Tables** page, choose a table in open table format that you want to enable compaction for, then under **Actions** menu, choose **Optimization**, and then choose **Enable**.

   You can also enable compaction by selecting the **Table optimization** tab on the **Table details** page. Choose the **Table optimization** tab on the lower section of the page, and choose **Enable compaction**.

The **Enable optimization** option is also available when you create a new Iceberg table in the Data Catalog.

4. On the **Enable optimization** page, choose **Compaction** under **Optimization options**.



Optimization configuration

Configure Apache Iceberg table optimization to optimize storage and improve query performances. Learn more

○ Use default settings
Default settings have been defined to help you get started. You can change them at any time later.

● Customize settings
Customize settings for your specific needs.

IAM Role

Configure for all selected optimizers.

☑ Apply the selected IAM role to all selected optimizers. Update individual roles later as needed.
Create a single role with necessary permissions to configure all table optimizers. Learn more

IAM role

Choose an IAM role

Virtual private cloud (VPC) - ***optional***

Choose a network connection to access a data store in your VPC.

Choose a Glue connection

Compaction configuration

Compaction strategy

Choose a compaction strategy Learn more

● Binpack
Combines small files into larger files, typically targeting sizes over 100MB, while applying any pending deletes. This is the default and recommended compaction strategy for most use cases.

○ Sort
Organizes data based on specified columns which are sorted hierarchically during compaction, improving query performance for filtered operations. This strategy is recommended when your queries frequently filter on specific columns. To use this strategy, you must first define a sort order in your Iceberg table properties using the 'sort_order' table property.

○ Z-Order
Optimizes data organization by blending multiple attributes into a single scalar value that can be used for sorting, allowing efficient querying across multiple dimensions. This strategy is recommended when you need to query data across multiple dimensions simultaneously. To use this strategy, you must first define a sort order in your Apache Iceberg table properties using the 'sort_order' table property.

Minimum input files

Minimum number of data-files to be present in a partition before compaction will actually compact files.

100 files

Default is 100 files.

Delete file threshold

Minimum number of deletes in a data-file to be eligible for compaction.

1 deletes

Default is 1 delete.

5. Next, select an IAM role from the drop down with the permissions shown in the [Table optimization prerequisites](#) section.

   You can also choose **Create a new IAM role** option to create a custom role with the required permissions to run compaction.

   Follow the steps below to update an existing IAM role:

   a. To update the permissions policy for the IAM role, in the IAM console, go to the IAM role that is being used for running compaction.

   b. In the **Add permissions** section, choose Create policy. In the newly opened browser window, create a new policy to use with your role.

   c. On the Create policy page, choose the `JSON` tab. Copy the JSON code shown in the Prerequisites into the policy editor field.

6. If you have security policy configurations where the Iceberg table optimizer needs to access Amazon S3 buckets from a specific Virtual Private Cloud (VPC), create an AWS Glue network connection or use an existing one.

   If you don't have an AWS Glue VPC connection set up already, create a new one by following the steps in the [Creating connections for connectors](#) section using the AWS Glue console or the AWS CLI/SDK.

7. Choose a compaction strategy. The available options are:

   - **Binpack** – Binpack is the default compaction strategy in Apache Iceberg. It combines smaller data files into larger ones for optimal performance.
   - **Sort** – Sorting in Apache Iceberg is a data organization technique that clusters information within files based on specified columns, significantly improving query performance by reducing the number of files that need to be processed. You define the sort order in Iceberg's metadata using the sort-order field, and when multiple columns are specified, data is sorted in the sequence the columns appear in the sort order, ensuring records with similar values are stored together within files. The sorting compaction strategy takes the optimization further by sorting data across all files within a partition.
   - **Z-order** – Z-ordering is a way to organize data when you need to sort by multiple columns with equal importance. Unlike traditional sorting that prioritizes one column over others, Z-ordering gives balanced weight to each column, helping your query engine read fewer files when searching for data.

The technique works by weaving together the binary digits of values from different columns. For example, if you have the numbers 3 and 4 from two columns, Z-ordering first converts them to binary (3 becomes 011 and 4 becomes 100), then interleaves these digits to create a new value: 011010. This interleaving creates a pattern that keeps related data physically close together.

Z-ordering is particularly effective for multi-dimensional queries. For example, a customer table Z-ordered by income, state, and zip code can deliver superior performance compared to hierarchical sorting when querying across multiple dimensions. This organization allows queries targeting specific combinations of income and geographic location to quickly locate relevant data while minimizing unnecessary file scans.

8. **Minimum input files** – The number of data files required in a partition before compaction is triggered.
9. **Delete files threshold** – Minimum delete operations required in a data file before it becomes eligible for compaction.
10. Choose **Enable optimization**.

AWS CLI

The following example shows how to enable compaction. Replace the account ID with a valid AWS account ID. Replace the database name and table name with actual Iceberg table name and the database name. Replace the `roleArn` with the AWS Resource Name (ARN) of the IAM role and name of the IAM role that has the required permissions to run compaction. You can replace compaction strategy `sort` with other supported strategies like `z-order` or `binpack`.

order" depending on your requirements.

```
aws glue create-table-optimizer \
  --catalog-id 123456789012 \
  --database-name iceberg_db \
  --table-name iceberg_table \
  --table-optimizer-configuration '{
    "roleArn": "arn:aws:iam::123456789012:role/optimizer_role",
    "enabled": true,
    "vpcConfiguration": {"glueConnectionName": "glue_connection_name"},
    "compactionConfiguration": {
      "icebergConfiguration": {"strategy": "sort"}
    }
  }'\
--type compaction
```

AWS API

Call CreateTableOptimizer operation to enable compaction for a table.

After you enable compaction, **Table optimization** tab shows the following compaction details once the compaction run is complete:

Start time

The time at which the compaction process started within Data Catalog. The value is a timestamp in UTC time.

End time

The time at which the compaction process ended in Data Catalog. The value is a timestamp in UTC time.

Status

The status of the compaction run. Values are success or fail.

Files compacted

Total number of files compacted.

Bytes compacted

Total number of bytes compacted.

Disabling compaction optimizer

You can disable automatic compaction for a particular Apache Iceberg table using AWS Glue console or AWS CLI.

Console

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. On the left navigation, under **Data Catalog**, choose **Tables**.
3. From the tables list, choose the Iceberg table that you want to disable compaction.

4. Choose the **Table optimization** tab on the lower section of the **Tables details** page.
5. From **Actions**, choose **Disable**, and then choose **Compaction**.
6. Choose **Disable compaction** on the confirmation message. You can re-enable compaction at a later time.

   After the you confirm, compaction is disabled and the compaction status for the table turns back to `Disabled`.

AWS CLI

In the following example, replace the account ID with a valid AWS account ID. Replace the database name and table name with actual Iceberg table name and the database name. Replace the `roleArn` with the AWS Resource Name (ARN) of the IAM role and actual name of the IAM role that has the required permissions to run compaction.

```
aws glue update-table-optimizer \
  --catalog-id 123456789012 \
  --database-name iceberg_db \
  --table-name iceberg_table \
  --table-optimizer-configuration
 '{"roleArn":"arn:aws:iam::123456789012:role/optimizer_role", "enabled":'false',
 "vpcConfiguration":{"glueConnectionName":"glue_connection_name"}}'\
  --type compaction
```

AWS API

Call UpdateTableOptimizer operation to disable compaction for a specific table.

Snapshot retention optimization

Apache Iceberg snapshot retention feature allows users to query historical data at specific points in time and revert unwanted modifications to their tables. In the AWS Glue Data Catalog, snapshot retention configuration controls how long these snapshots (versions of the table data) are kept before being expired and removed. This helps manage storage costs and metadata overhead by automatically removing older snapshots based on a configured retention period or maximum number of snapshots to keep.

You can configure the retention period in days and the maximum number of snapshots to retain for a table. AWS Glue removes snapshots that are older than the specified retention period from the table metadata, while keeping the most recent snapshots up to the configured limit. After removing old snapshots from the metadata, AWS Glue deletes the corresponding data and metadata files that are no longer referenced and unique to the expired snapshots. This allows time travel queries only up to the remaining retained snapshots, while reclaiming storage space used by expired snapshot data.

Topics

- Enabling snapshot retention optimizer
- Updating snapshot retention optimizer
- Disabling snapshot retention optimizer

Enabling snapshot retention optimizer

You can use AWS Glue console, AWS CLI, or AWS API to enable snapshot retention optimizers for your Apache Iceberg tables in the Data Catalog. For new tables, you can choose Apache Iceberg as table format and enable snapshot retention optimizer when you create the table. Snapshot retention is disabled by default for new tables.

Console

To enable snapshot retention optimizer

1. Open the AWS Glue console at https://console.aws.amazon.com/glue/ and sign in as a data lake administrator, the table creator, or a user who has been granted the `glue:UpdateTable` and `lakeformation:GetDataAccess` permissions on the table.
2. In the navigation pane, under **Data Catalog**, choose **Tables**.
3. On the **Tables** page, choose an Iceberg table that you want to enable snapshot retention optimizer for, then under **Actions** menu, choose **Enable** under **Optimization**.

   You can also enable optimization by selecting the table and opening the **Table details** page. Choose the **Table optimization** tab on the lower section of the page, and choose **Enable snapshot retention**.
4. On the **Enable optimization** page, under **Optimization configuration**, you have two options: **Use default setting** or **Customize settings**. If you choose to use the default settings, AWS Glue utilizes the properties defined in the Iceberg table configuration to determine the snapshot retention period and the number of snapshots to be retained. In the absence of this

configuration, AWS Glue retains one snapshot for five days, and deletes files associated with the expired snapshots.

5. Next, choose an IAM role that AWS Glue can assume on your behalf to run the optimizer. For details about the permissions required for the IAM role, see the [Table optimization prerequisites](#) section.

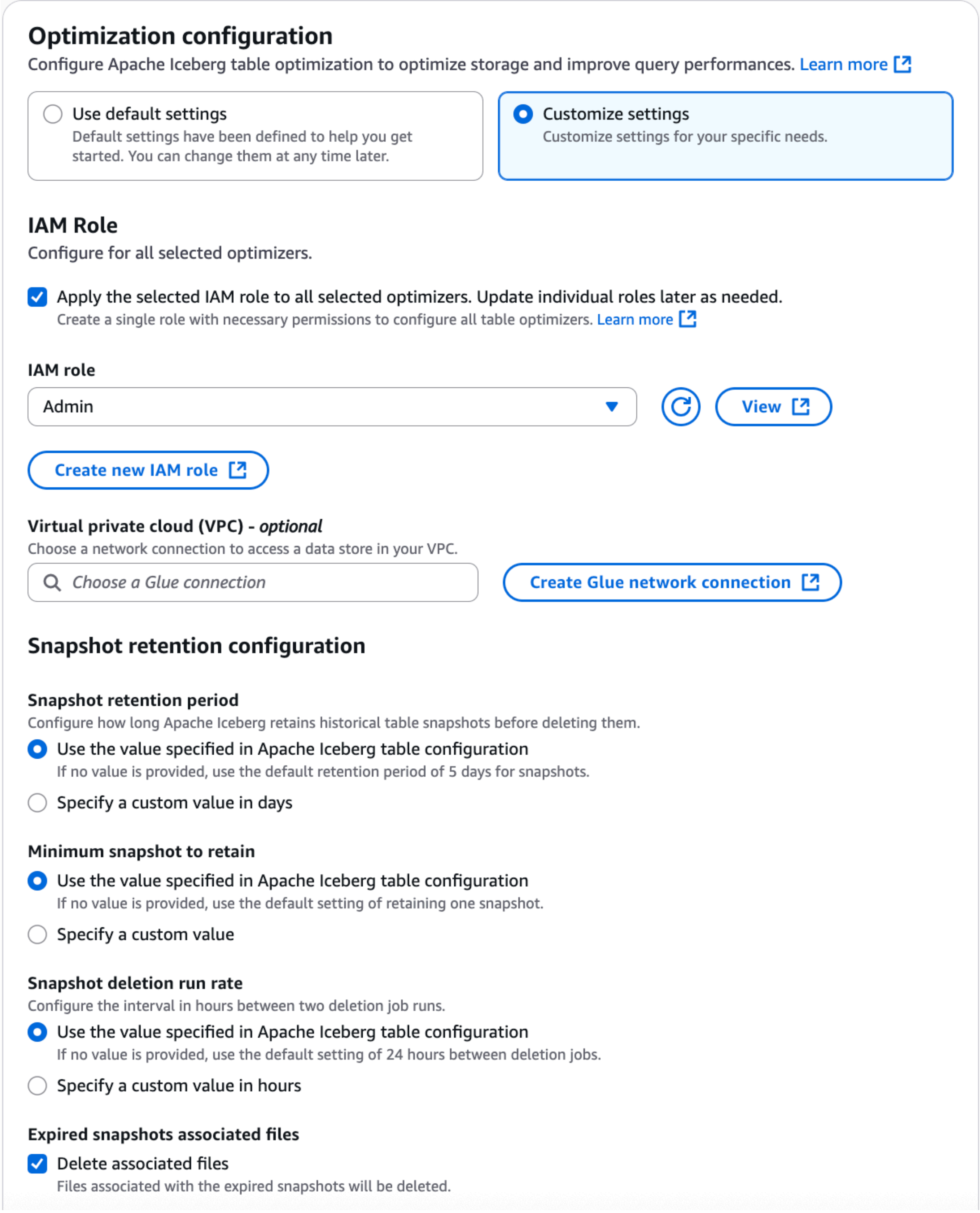
   Follow the steps below to update an existing IAM role:

   a. To update the permissions policy for the IAM role, in the IAM console, go to the IAM role that is being used for running compaction.

   b. In the Add permissions section, choose Create policy. In the newly opened browser window, create a new policy to use with your role.

   c. On the Create policy page, choose the JSON tab. Copy the JSON code shown in the Prerequisites into the policy editor field.

6. If you prefer to set the values for the **Snapshot retention configuration** manually, choose **Customize settings**.



Optimization configuration

Configure Apache Iceberg table optimization to optimize storage and improve query performances. Learn more

○ Use default settings
Default settings have been defined to help you get started. You can change them at any time later.

● Customize settings
Customize settings for your specific needs.

IAM Role

Configure for all selected optimizers.

☑ Apply the selected IAM role to all selected optimizers. Update individual roles later as needed.
Create a single role with necessary permissions to configure all table optimizers. Learn more

IAM role

Admin

Virtual private cloud (VPC) - ***optional***

Choose a network connection to access a data store in your VPC.

Snapshot retention configuration

Snapshot retention period

Configure how long Apache Iceberg retains historical table snapshots before deleting them.

● Use the value specified in Apache Iceberg table configuration
If no value is provided, use the default retention period of 5 days for snapshots.

○ Specify a custom value in days

Minimum snapshot to retain

● Use the value specified in Apache Iceberg table configuration
If no value is provided, use the default setting of retaining one snapshot.

○ Specify a custom value

Snapshot deletion run rate

Configure the interval in hours between two deletion job runs.

● Use the value specified in Apache Iceberg table configuration
If no value is provided, use the default setting of 24 hours between deletion jobs.

○ Specify a custom value in hours

Expired snapshots associated files

☑ Delete associated files
Files associated with the expired snapshots will be deleted.

7. Choose the box **Apply the selected IAM role to the selected optimizers** option to use a single IAM role for all enabling all optimizers.
8. If you have security policy configurations where the Iceberg table optimizer needs to access Amazon S3 buckets from a specific Virtual Private Cloud (VPC), create an AWS Glue network connection or use an existing one.

   If you don't have an AWS Glue VPC Connection set up already, create a new one by following the steps in the Creating connections for connectors section using the AWS Glue console or the AWS CLI/SDK.
9. Next, under **Snapshot retention configuration**, either choose to use the values specified in the Iceberg table configuration, or specify custom values for snapshot retention period (history.expire.max-snapshot-age-ms), minimum number of snapshots (history.expire.min-snapshots-to-keep) to retain, and the time in hours between consecutive snapshot deletion job runs.
10. Choose **Delete associated files** to delete underlying files when the table optimizer deletes old snapshots from the table metadata.

    If you don't choose this option, when older snapshots are removed from the table metadata, their associated files will remain in the storage as orphaned files.
11. Next, read the caution statement, and choose **I acknowledge** to proceed.

    > **Note**
    >
    > In the Data Catalog, the snapshot retention optimizer honors the lifecycle that is controlled by branch and tag level retention policies. For more information, see Branching and tagging section in the Iceberg documentation.
12. Review the configuration and choose **Enable optimization**.

    Wait a few minutes for the retention optimizer to run and expire old snapshots based on the configuration.

AWS CLI

To enable snapshot retention for new Iceberg tables in AWS Glue, you need to create a table optimizer of type `retention` and set the `enabled` field to `true` in the `table-optimizer-configuration`. You can do this using the AWS CLI command `create-table-optimizer` or `update-table-optimizer`. Additionally, you need to specify the retention configuration fields like `snapshotRetentionPeriodInDays` and `numberOfSnapshotsToRetain` based on your requirements.

The following example shows how to enable the snapshot retention optimizer. Replace the account ID with a valid AWS account ID. Replace the database name and table name with actual Iceberg table name and the database name. Replace the `roleArn` with the AWS Resource Name (ARN) of the IAM role and name of the IAM role that has the required permissions to run the snapshot retention optimizer.

```
aws glue create-table-optimizer \
  --catalog-id 123456789012 \
  --database-name iceberg_db \
  --table-name iceberg_table \
  --table-optimizer-configuration
 '{"roleArn":"arn:aws:iam::123456789012:role/optimizer_role","enabled":'true',
 "vpcConfiguration":{
"glueConnectionName":"glue_connection_name"}, "retentionConfiguration":
{"icebergConfiguration":
{"snapshotRetentionPeriodInDays":7,"numberOfSnapshotsToRetain":3,"cleanExpiredFiles":'true'}
  --type retention
```

This command creates a retention optimizer for the specified Iceberg table in the given catalog, database, and Region. The table-optimizer-configuration specifies the IAM role ARN to use, enables the optimizer, and sets the retention configuration. In this example, it retains snapshots for 7 days, keeps a minimum of 3 snapshots, and cleans expired files.

- snapshotRetentionPeriodInDays –The number of days to retain snapshots before expiring them. The default value is 5.
- numberOfSnapshotsToRetain – The minimum number of snapshots to keep, even if they are older than the retention period. The default value is 1.
- cleanExpiredFiles – A boolean indicating whether to delete expired data files after expiring snapshots. The default value is `true`.

  When set to true, older snapshots are removed from table metadata, and their underlying files are deleted. If this parameter is set to false, older snapshots are removed from table metadata but their underlying files remain in the storage as orphan files.

AWS API

Call CreateTableOptimizer operation to enable snapshot retention optimizer for a table.

After you enable compaction, **Table optimization** tab shows the following compaction details (after approximately 15-20 minutes):

Start time

The time at which the snapshot retention optimizer started. The value is a timestamp in UTC time.

Run time

The time shows how long the optimizer takes to complete the task. The value is a timestamp in UTC time.

Status

The status of the optimizer run. Values are success or fail.

Data files deleted

Total number of files deleted.

Manifest files deleted

Total number of manifest files deleted.

Manifest lists deleted

Total number of manifest lists deleted.

Updating snapshot retention optimizer

You can update the existing configuration of an snapshot retention optimizer for a particular Apache Iceberg table using the AWS Glue console, AWS CLI, or the UpdateTableOptimizer API.

Console

To update snapshot retention configuration

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.

2. Choose **Data Catalog** and choose **Tables**. From the tables list, choose the Iceberg table you want to update the snapshot retention optimizer configuration.
3. On the lower section of the **Tables details** page, select the **Table optimization** tab, and then choose **Edit**. You can also choose **Edit** under **Optimization** from the **Actions** menu located on the top right corner of the page.
4. On the **Edit optimization** page, make the desired changes.
5. Choose **Save**.

AWS CLI

To update a snapshot retention optimizer using the AWS CLI, you can use the following command:

```
aws glue update-table-optimizer \
 --catalog-id 123456789012 \
 --database-name iceberg_db \
 --table-name iceberg_table \
 --table-optimizer-configuration
 '{"roleArn":"arn:aws:iam::123456789012:role/optimizer_role"","enabled":'true',
 "vpcConfiguration":
{"glueConnectionName":"glue_connection_name"},"retentionConfiguration":
{"icebergConfiguration":
{"snapshotRetentionPeriodInDays":7,"numberOfSnapshotsToRetain":3,"cleanExpiredFiles":'true'}
 \
 --type retention
```

This command updates the retention configuration for the specified table in the given catalog, database, and Region. The key parameters are:

- snapshotRetentionPeriodInDays –The number of days to retain snapshots before expiring them. The default value is 1.
- numberOfSnapshotsToRetain – The minimum number of snapshots to keep, even if they are older than the retention period. The default value is 5.
- cleanExpiredFiles – A boolean indicating whether to delete expired data files after expiring snapshots. The default value is `true`.

When set to true, older snapshots are removed from table metadata, and their underlying files are deleted." If this parameter is set to false, older snapshots are removed from table metadata but their underlying files remain in the storage as orphan files.

API

To update a table optimizer, you can use the `UpdateTableOptimizer` API. This API allows you to update the configuration of an existing table optimizer for compaction, retention, or orphan file removal. The request parameters include:

- catalogId (required): The ID of the catalog containing the table
- databaseName (optional): The name of the database containing the table
- tableName (optional): The name of the table
- type (required): The type of table optimizer (compaction, retention, or orphan_file_deletion)
- retentionConfiguration (required): The updated configuration for the table optimizer, including role ARN, enabled status, retention configuration, and orphan file removal configuration.

Disabling snapshot retention optimizer

You can disable the snapshot retention optimizer for a particular Apache Iceberg table using AWS Glue console or AWS CLI.

Console

To disable snapshot retention

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. Choose **Data Catalog** and choose **Tables**. From the tables list, choose the Iceberg table that you want to disable the optimizer for snapshot retention.
3. On lower section of the **Table details** page, choose **Table optimization** and **Disable**, **Snapshot retention** under **Actions**.

   You can also choose **Disable** under **Optimization** from the **Actions** menu located on top right corner of the page.

4. Choose **Disable** on the confirmation message. You can re-enable the snapshot retention optimizer at a later time.

   After the you confirm, snapshot retention optimizer is disabled and the status for snapshot retention turns back to `Not enabled`.

AWS CLI

In the following example, replace the account ID with a valid AWS account ID. Replace the database name and table name with actual Iceberg table name and the database name. Replace the `roleArn` with the AWS Resource Name (ARN) of the IAM role and actual name of the IAM role that has the required permissions to run the retention optimizer.

```
aws glue update-table-optimizer \
  --catalog-id 123456789012 \
  --database-name iceberg_db \
  --table-name iceberg_table \
  --table-optimizer-configuration
 '{"roleArn":"arn:aws:iam::123456789012:role/optimizer_role", "vpcConfiguration":
{"glueConnectionName":"glue_connection_name"}, "enabled":'false'}'\
  --type retention
```

AWS API

Call UpdateTableOptimizer operation to disable the snapshot retention optimizer for a specific table.

Deleting orphan files

AWS Glue Data Catalog allows you to remove orphan files from your Iceberg tables. Orphan files are unreferenced files that exist in your Amazon S3 data source under the specified table location, are not tracked by the Iceberg table metadata, and are older than your configured age limit. These orphan files can accumulate over time due to failure in operations like compaction, partition drops, or table rewrites, and take up unnecessary storage space.

The orphan file deletion optimizer in AWS Glue scans the table metadata and the actual data files, identifies the orphan files, and deletes them to reclaim storage space. The optimizer only removes files created after the optimizer's creation date that also meet the configured deletion criteria. Files created before or on the optimizer creation date are never deleted.

Orphan file deletion logic

1. Date check – Compares file creation date with optimizer creation date. If file is older than or equal to optimizer creation date, the file is skipped.
2. Optimizer configuration check – If file is newer than optimizer creation date, evaluates the file against the configured age limit. The optimizer deletes the file if it matches the deletion critera. Skips the file, if it doesn't match the criteria.

You can initiate the orphan file deletion by creating an orphan file deletion table optimizer in the Data Catalog.

> ⚠ **Important**
>
> By default, orphan file deletion evaluates files across your AWS Glue table location. While you can configure a sub-prefix to limit the scope of evaluation by using API parameter, you must ensure your table location doesn't contain files from other data sources or tables. If your table location overlaps with other data sources, the service might identify and delete unrelated files as orphans.

Topics

- Enabling orphan file deletion
- Updating orphan file deletion optimizer
- Disabling orphan file deletion

Enabling orphan file deletion

You can use AWS Glue console, AWS CLI, or AWS API to enable orphan file deletion for your Apache Iceberg tables in the Data Catalog. For new tables, you can choose Apache Iceberg as table format and enable orphan file deletion optimizer when you create the table. Snapshot retention is disabled by default for new tables.

Console

To enable orphan file deletion

1. Open the AWS Glue console at https://console.aws.amazon.com/glue/ and sign in as a data lake administrator, the table creator, or a user who has been granted the `glue:UpdateTable` and `lakeformation:GetDataAccess` permissions on the table.
2. In the navigation pane, under **Data Catalog**, choose **Tables**.
3. On the **Tables** page, choose an Iceberg table in that you want to enable orphan file deletion.

   Choose the **Table optimization** tab on the lower section of the page, and choose **Enable**, **Orphan file deletion** from **Actions**.

   You can also choose **Enable** under **Optimization** from the **Actions** menu located on the top right corner of the page..
4. On the **Enable optimization** page, choose **Orphan file deletion** under **Optimization options**.
5. If you choose to use **Default settings**, all orphan files will be deleted after 3 days. If you want to keep the orphan files for a specific number of days, choose **Customize settings**.
6. Next, choose an IAM role with the required permissions to delete orphan files.
7. If you have security policy configurations where the Iceberg table optimizer needs to access Amazon S3 buckets from a specific Virtual Private Cloud (VPC), create an AWS Glue network connection or use an existing one.

   If you don't have an AWS Glue VPC Connection set up already, create a new one by following the steps in the Creating connections for connectors section using the AWS Glue console or the AWS CLI/SDK.
8. If you choose **Customize settings**, enter the number of days to retain the files before deletion under **Orphan file deletion configuration**. You can also specify the interval between two consecutive optimizer runs. The default value is 24 hours.
9. Choose **Enable optimization**.

AWS CLI

To enable orphan file deletion for an Iceberg table in AWS Glue, you need to create a table optimizer of type `orphan_file_deletion` and set the `enabled` field to true. To create an orphan file deletion optimizer for an Iceberg table using the AWS CLI, you can use the following command:

```
aws glue create-table-optimizer \
 --catalog-id 123456789012 \
 --database-name iceberg_db \
 --table-name iceberg_table \
 --table-optimizer-configuration
 '{"roleArn":"arn:aws:iam::123456789012:role/optimizer_role","enabled":true,
 "vpcConfiguration":{
"glueConnectionName":"glue_connection_name"}, "orphanFileDeletionConfiguration":
{"icebergConfiguration":{"orphanFileRetentionPeriodInDays":3, "location":'S3
 location'}}}'\
 --type orphan_file_deletion
```

This command creates an orphan file deletion optimizer for the specified Iceberg table. The key parameters are:

- roleArn – the ARN of the IAM role with permissions to access the S3 bucket and Glue resources.
- enabled – Set to true to enable the optimizer.
- orphanFileRetentionPeriodInDays – The number of days to retain orphan files before deleting them (minimum 1 day).
- type – Set to orphan_file_deletion to create an orphan file deletion optimizer.

After creating the table optimizer, it will run orphan file deletion periodically (once per day if left enabled). You can check the runs using the `list-table-optimizer-runs` API. The orphan file deletion job will identify and delete files that are not tracked in the Iceberg metadata for the table.

API

Call CreateTableOptimizer operation to create the orphan file deletion optimizer for a specific table.

Updating orphan file deletion optimizer

You can modify the configuration for the orphan file deletion optimizer, such as changing the retention period for orphan files or the IAM role used by the optimizer using AWS Glue console, AWS CLI, or the `UpdateTableOptimizer` operation.

AWS Management Console

To update the orphan file deletion optimizer

1. Choose **Data Catalog** and choose **Tables**. From the tables list, choose the table you want to update the orphan file deletion optimizer configuration.
2. On the lower section of the **Tables details** page, choose **Table optimization** , and then choose **Edit**.
3. On the **Edit optimization** page, make the desired changes.
4. Choose **Save**.

AWS CLI

You can use the `update-table-optimizer` call to update the orphan file deletion optimizer in AWS Glue, you can use call. This allows you to modify the `OrphanFileDeletionConfiguration` in the `icebergConfiguration` field where you can specify the updated `OrphanFileRetentionPeriodInDays` to set the number of days to retain orphan files, to specify the Iceberg table location to delete orphan files from.

```
aws glue update-table-optimizer \
 --catalog-id 123456789012 \
 --database-name iceberg_db \
 --table-name Iceberg_table \
 --table-optimizer-configuration
 '{"roleArn":"arn:aws:iam::123456789012:role/optimizer_role","enabled":true,
 "vpcConfiguration":
{"glueConnectionName":"glue_connection_name"},"orphanFileDeletionConfiguration":
{"icebergConfiguration":{"orphanFileRetentionPeriodInDays":5}}}' \
 --type orphan_file_deletion
```

API

Call the [UpdateTableOptimizer]() operation to update the orphan file deletion optimizer for a table.

Disabling orphan file deletion

You can disable orphan file deletion optimizer for a particular Apache Iceberg table using AWS Glue console or AWS CLI.

Console

To disable orphan file deletion

1. Choose **Data Catalog** and choose **Tables**. From the tables list, choose the Iceberg table that you want to disable the optimizer for orphan file deletion.
2. On lower section of the **Table details** page, choose **Table optimization** tab.
3. Choose **Actions**, and then choose **Disable** , **Orphan file deletion**.

   You can also choose **Disable** under **Optimization** from the **Actions** menu.
4. Choose **Disable** on the confirmation message. You can re-enable the orphan file deletion optimizer at a later time.

   After the you confirm, orphan file deletion optimizer is disabled and the status for orphan file deletion turns back to `Not enabled`.

AWS CLI

In the following example, replace the account ID with a valid AWS account ID. Replace the database name and table name with actual Iceberg table name and the database name. Replace the `roleArn` with the AWS Resource Name (ARN) of the IAM role and actual name of the IAM role that has the required permissions to disable the optimizer.

```
aws glue update-table-optimizer \
  --catalog-id 123456789012 \
  --database-name iceberg_db \
  --table-name iceberg_table \
  --table-optimizer-configuration
 '{"roleArn":"arn:aws:iam::123456789012:role/optimizer_role", "enabled":'false'}'\
  --type orphan_file_deletion
```

API

Call the UpdateTableOptimizer operation to disable the snapshot retention optimizer for a specific table.

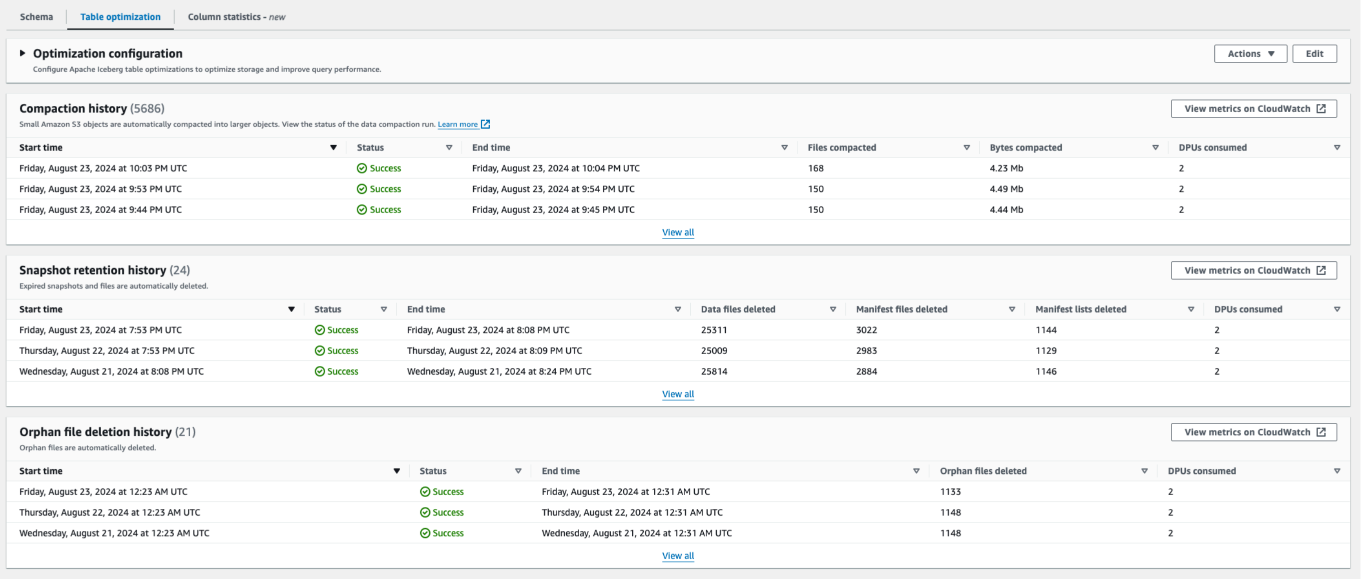
Viewing optimization details

You can view the optimization status for Apache Iceberg tables in the AWS Glue console, AWS CLI, or using AWS API operations.

Console

To view the optimization status for Iceberg tables (console)

- You can view optimization status for Iceberg tables on the AWS Glue console by choosing an Iceberg table from the **Tables** list under **Data Catalog**. Under **Table optimization**. Choose the **View all**



Optimization configuration

Configure Apache Iceberg table optimizations to optimize storage and improve query performance.

Compaction history (5686)

Small Amazon S3 objects are automatically compacted into larger objects. View the status of the data compaction run.

Start time	Status	End time	Files compacted	Bytes compacted	DPUs consumed
Friday, August 23, 2024 at 10:03 PM UTC	✓ Success	Friday, August 23, 2024 at 10:04 PM UTC	168	4.23 Mb	2
Friday, August 23, 2024 at 9:53 PM UTC	✓ Success	Friday, August 23, 2024 at 9:54 PM UTC	150	4.49 Mb	2
Friday, August 23, 2024 at 9:44 PM UTC	✓ Success	Friday, August 23, 2024 at 9:45 PM UTC	150	4.44 Mb	2

Snapshot retention history (24)

Expired snapshots and files are automatically deleted.

Start time	Status	End time	Data files deleted	Manifest files deleted	Manifest lists deleted	DPUs consumed
Friday, August 23, 2024 at 7:53 PM UTC	✓ Success	Friday, August 23, 2024 at 8:08 PM UTC	25311	3022	1144	2
Thursday, August 22, 2024 at 7:53 PM UTC	✓ Success	Thursday, August 22, 2024 at 8:09 PM UTC	25009	2983	1129	2
Wednesday, August 21, 2024 at 8:08 PM UTC	✓ Success	Wednesday, August 21, 2024 at 8:24 PM UTC	25814	2884	1146	2

Orphan file deletion history (21)

Orphan files are automatically deleted.

Start time	Status	End time	Orphan files deleted	DPUs consumed
Friday, August 23, 2024 at 12:23 AM UTC	✓ Success	Friday, August 23, 2024 at 12:31 AM UTC	1133	2
Thursday, August 22, 2024 at 12:23 AM UTC	✓ Success	Thursday, August 22, 2024 at 12:31 AM UTC	1148	2
Wednesday, August 21, 2024 at 12:23 AM UTC	✓ Success	Wednesday, August 21, 2024 at 12:31 AM UTC	1148	2

AWS CLI

You can view the optimization details using AWS CLI.

In the following examples, replace the account ID with a valid AWS account ID, the database name, and table name with actual Iceberg table name. For `type`, provide and optimization type. Acceptable values are `compaction`, `retention`, and `orphan-file-deletion`.

- **To get the last compaction run details for a table**

```
aws get-table-optimizer \
  --catalog-id 123456789012 \
  --database-name iceberg_db \
  --table-name iceberg_table \
  --type compaction
```

- Use the following example to retrieve the history of an optimizer for a specific table.

```
aws list-table-optimizer-runs \
  --catalog-id 123456789012 \
  --database-name iceberg_db \
  --table-name iceberg_table \
  --type compaction
```

- The following example shows how to retrieve the optimization run and configuration details for multiple optimizers. You can specify a maximum of 20 optimizers.

```
aws glue batch-get-table-optimizer \
--entries '[{"catalogId":"123456789012", "databaseName":"iceberg_db",
 "tableName":"iceberg_table", "type":"compaction"}]'
```

API

- Use `GetTableOptimizer` operation to retrieve the last run details of an optimizer.
- Use `ListTableOptimizerRuns` operation to retrieve history of a given optimizer on a specific table. You can specify 20 optimizers in a single API call.
- Use the BatchGetTableOptimizer operation to retrieve configuration details for multiple optimizers in your account.

Viewing Amazon CloudWatch metrics

After running the table optimizers successfully, the service creates Amazon CloudWatch metrics on the optimization job performance. You can go to the **CloudWatch Metrics** and choose **Metrics**, **All metrics**. You can to filter metrics by the specific namespace (for example AWS Glue), table name, or database name.

For more information, see View available metrics in the *Amazon CloudWatch User Guide*.

Compaction

- Number of bytes compacted
- Number of files compacted

- Number of DPU allocated to job
- Duration of job (Hours)

Snapshot retention

- Number of data files deleted
- Number of manifest files deleted
- Number of Manifest lists deleted
- Duration of job (Hours)

Orphan file deletion

- Number of orphan files deleted
- Duration of job (Hours)

Deleting an optimizer

You can delete an optimizer and associated metadata for the table using AWS CLI or AWS API operation.

Run the following AWS CLI command to delete optimization history for a table. You need to specify the optimizer `type` along with the catalog ID, database name and table name. The acceptable values are: `compaction`, `retention`, and `orphan_file_deletion`.

```
aws glue delete-table-optimizer \
  --catalog-id 123456789012 \
  --database-name iceberg_db \
  --table-name iceberg_table \
  --type compaction
```

Use `DeleteTableOptimizer` operation to delete an optimizer for a table.

Considerations and limitations

This section includes things to consider when using table optimizers within the AWS Glue Data Catalog.

Supported formats and limitations for managed data compaction

Data compaction supports a variety of data types and compression formats for reading and writing data, including reading data from encrypted tables.

Data compaction supports:

- **Encryption –** Data compaction only supports default Amazon S3 encryption (SSE-S3) and server-side KMS encryption (SSE-KMS).
- **Compaction strategies** –- Binpack, sort, and Z-order sorting
- You can run compaction from the account where Data Catalog resides when the Amazon S3 bucket that stores the underlying data is in another account. To do this, the compaction role requires access to the Amazon S3 bucket.

Data compaction currently doesn’t support:

- **Compaction on cross-account tables –** You can't run compaction on cross-account tables.
- **Compaction on cross-Region tables –** You can't run compaction on cross-Region tables.
- **Enabling compaction on resource links**
- **Tables in Amazon S3 Express One Zone storage class** – You can't run compaction on S3 Express One Zone Iceberg Tables.
- **Z-order compaction strategy doesn't support the following data types :**
  - Decimal
  - TimestampWithoutZone

Considerations for snapshot retention and orphan file deletion optimizers

The following considerations apply to the snapshot retention and the orphan file deletion optimizers.

- The snapshot retention and orphan file deletion processes have a maximum limit of deleting 1,000,000 files per run. When deleting expired snapshots, if the number of eligible files for deletion surpasses 1,000,000, any remaining files beyond that threshold will continue to exist in the table storage as orphan files.
- Snapshots will be preserved by the snapshot retention optimizer only when both criteria are satisfied: the minimum number of snapshots to keep and the specified retention period.

- The snapshot retention optimizer deletes expired snapshot metadata from Apache Iceberg, preventing time travel queries for expired snapshots and optionally deleting associated data files.
- Orphan file deletion optimizer deletes orphaned data and metadata files that are no longer referenced by Iceberg metadata if their creation time is before the orphan file deletion retention period from the time of optimizer run.
- Apache Iceberg facilitates version control through branches and tags, which are named pointers to specific snapshot states. Each branch and tag follows its own independent life-cycle, governed by retention policies defined at their respective levels. The AWS Glue Data Catalog optimizers take these life cycle policies into account, ensuring adherence to the specified retention rules. Branch and tag-level retention policies take precedence over the optimizer configurations.

  For more information, see Branching and Tagging in Apache Iceberg documentation.
- Snapshot retention and orphan file deletion optimizers will delete files eligible for clean-up as per configured parameters. Enhance your control over file deletion by implementing S3 versioning and life-cycle policies on the appropriate buckets.

  For detailed instructions on setting up versioning and creating life cycle rules, see https://docs.aws.amazon.com/AmazonS3/latest/userguide/Versioning.html.
- For proper orphan file determination, ensure that the provided table location and any sub-paths don't overlap with or contain data from any other tables or data sources. If paths overlap, you risk unrecoverable data loss from unintended deletion of files.

Supported Regions for table optimizers

The table optimization features (compaction, snapshot retention, and orphan file deletion) for AWS Glue Data Catalog are available in the following AWS Regions:

- Asia Pacific (Tokyo)
- Asia Pacific (Seoul)
- Asia Pacific (Mumbai)
- Asia Pacific (Singapore)
- Asia Pacific (Sydney)
- Asia Pacific (Jakarta)
- Canada (Central)

- Europe (Ireland)
- Europe (London)
- Europe (Frankfurt)
- Europe (Stockholm)
- US East (N. Virginia)
- US East (Ohio)
- US West (Oregon)
- South America (São Paulo)

Optimizing query performance for Iceberg tables

Apache Iceberg is a high-performance open table format for huge analytic datasets. AWS Glue supports calculating and updating number of distinct values (NDVs) for each column in Iceberg tables. These statistics can facilitate better query optimization, data management, and performance efficiency for data engineers and scientists working with large-scale datasets.

AWS Glue estimates the number of distinct values in each column of the Iceberg table and and store them in Puffin files on Amazon S3 associated with Iceberg table snapshots. Puffin is an Iceberg file format designed to store metadata like indexes, statistics, and sketches. Storing sketches in Puffin files tied to snapshots ensures transactional consistency and freshness of the NDV statistics.

You can configure to run column statistics generation task using AWS Glue console or AWS CLI. When you initiate the process, AWS Glue starts a Spark job in the background and updates the AWS Glue table metadata in the Data Catalog. You can view column statistics using AWS Glue console or AWS CLI or by calling the GetColumnStatisticsForTable API operation.

> **Note**
>
> If you're using AWS Lake Formation permissions to control access to the table, the role assumed by the column statistics task requires full table access to generate statistics.

Topics

- Prerequisites for generating column statistics

- Generating column statistics for Iceberg tables
- See also

Prerequisites for generating column statistics

To generate or update column statistics for Iceberg tables, the statistics generation task assumes an AWS Identity and Access Management (IAM) role on your behalf. Based on the permissions granted to the role, the column statistics generation task can read the data from the Amazon S3 data store.

When you configure the column statistics generation task, AWS Glue allows you to create a role that includes the `AWSGlueServiceRole` AWS managed policy plus the required inline policy for the specified data source.

If you specify an existing role for generating column statistics, ensure that it includes the `AWSGlueServiceRole` policy or equivalent (or a scoped down version of this policy), and the required inline policies.

For more information about the required permissions, see Prerequisites for generating column statistics.

Generating column statistics for Iceberg tables

Follow these steps to configure a schedule for generating statistics in the Data Catalog using AWS Glue console or AWS CLI or the or run the **StartColumnStatisticsTaskRun** operation.

To generate column statistics

1. Sign in to the AWS Glue console at https://console.aws.amazon.com/glue/.
2. Choose **Tables** under Data Catalog .
3. Choose an Iceberg table from the list.
4. Choose **Column statistics**, **Generate on demand**,under **Actions** menu.

   You can also choose **Generate statistics** button under **Column statistics** tab in the lower section of the **Tables** page.
5. On the **Generate statistics** page, provide the statistics generation details. Follow steps 6-11 in the Generating column statistics on a schedule section to configure a schedule for statistics generation for Iceberg tables.

You can also choose to generate column statistics on-demand by followin the instructions in the Generating column statistics on demand

Note

Sampling option is not available for Iceberg tables.

AWS Glue calculates the number of distinct values for each column of the Iceberg table to a new Puffin file committed to the specified snapshot ID in your Amazon S3 location.

See also

- Viewing column statistics
- Viewing column statistics task runs
- Stopping column statistics task run
- Deleting column statistics

Managing the Data Catalog

The AWS Glue Data Catalog is a central metadata repository that stores structural and operational metadata for your Amazon S3 data sets. Managing the Data Catalog effectively is crucial for maintaining data quality, performance, security, and governance.

By understanding and applying these Data Catalog management practices, you can ensure your metadata remains accurate, performant, secure, and well-governed as your data landscape evolves.

This section covers the following aspects of Data Catalog management:

- *Updating table schema and partitions* As your data evolves, you may need to update the table schema or partition structure defined in the Data Catalog. For more information on how to make these updates programmatically using the AWS Glue ETL, see Updating the schema, and adding new partitions in the Data Catalog using AWS Glue ETL jobs.
- *Managing column statistics*: Accurate column statistics help optimize query plans and improve performance. For more information on how to generate, update, and manage column statistics, see Optimizing query performance using column statistics.

- *Encrypting the Data Catalog* To protect sensitive metadata, you can encrypt your Data Catalog using AWS Key Management Service (AWS KMS). This section explains how to enable and manage encryption for your Data Catalog.
- *Securing the Data Catalog with AWS Lake Formation* Lake Formation provides a comprehensive approach to data lake security and access control. You can use Lake Formation to secure and govern access to your Data Catalog and underlying data.

Topics

- Updating the schema, and adding new partitions in the Data Catalog using AWS Glue ETL jobs
- Optimizing query performance using column statistics
- Encrypting your Data Catalog
- Securing your Data Catalog using Lake Formation
- Working with AWS Glue Data Catalog views in AWS Glue

Updating the schema, and adding new partitions in the Data Catalog using AWS Glue ETL jobs

Your extract, transform, and load (ETL) job might create new table partitions in the target data store. Your dataset schema can evolve and diverge from the AWS Glue Data Catalog schema over time. AWS Glue ETL jobs now provide several features that you can use within your ETL script to update your schema and partitions in the Data Catalog. These features allow you to see the results of your ETL work in the Data Catalog, without having to rerun the crawler.

New partitions

If you want to view the new partitions in the AWS Glue Data Catalog, you can do one of the following:

- When the job finishes, rerun the crawler, and view the new partitions on the console when the crawler finishes.
- When the job finishes, view the new partitions on the console right away, without having to rerun the crawler. You can enable this feature by adding a few lines of code to your ETL script, as shown in the following examples. The code uses the `enableUpdateCatalog` argument to indicate that the Data Catalog is to be updated during the job run as the new partitions are created.

Method 1

Pass `enableUpdateCatalog` and `partitionKeys` in an options argument.

Python

```
additionalOptions = {"enableUpdateCatalog": True}
additionalOptions["partitionKeys"] = ["region", "year", "month", "day"]


sink = glueContext.write_dynamic_frame_from_catalog(frame=last_transform,
 database=<target_db_name>,

 table_name=<target_table_name>, transformation_ctx="write_sink",

 additional_options=additionalOptions)
```

Scala

```
val options = JsonOptions(Map(
    "path" -> <S3_output_path>,
    "partitionKeys" -> Seq("region", "year", "month", "day"),
    "enableUpdateCatalog" -> true))
val sink = glueContext.getCatalogSink(
    database = <target_db_name>,
    tableName = <target_table_name>,
    additionalOptions = options)sink.writeDynamicFrame(df)
```

Method 2

Pass `enableUpdateCatalog` and `partitionKeys` in `getSink()`, and call `setCatalogInfo()` on the `DataSink` object.

Python

```
sink = glueContext.getSink(
    connection_type="s3",
    path="<S3_output_path>",
    enableUpdateCatalog=True,
    partitionKeys=["region", "year", "month", "day"])
sink.setFormat("json")
sink.setCatalogInfo(catalogDatabase=<target_db_name>,
 catalogTableName=<target_table_name>)
```

```
sink.writeFrame(last_transform)
```

Scala

```
val options = JsonOptions(
   Map("path" -> <S3_output_path>,
       "partitionKeys" -> Seq("region", "year", "month", "day"),
       "enableUpdateCatalog" -> true))
val sink = glueContext.getSink("s3", options).withFormat("json")
sink.setCatalogInfo(<target_db_name>, <target_table_name>)
sink.writeDynamicFrame(df)
```

Now, you can create new catalog tables, update existing tables with modified schema, and add new table partitions in the Data Catalog using an AWS Glue ETL job itself, without the need to re-run crawlers.

Updating table schema

If you want to overwrite the Data Catalog table's schema you can do one of the following:

- When the job finishes, rerun the crawler and make sure your crawler is configured to update the table definition as well. View the new partitions on the console along with any schema updates, when the crawler finishes. For more information, see Configuring a Crawler Using the API.
- When the job finishes, view the modified schema on the console right away, without having to rerun the crawler. You can enable this feature by adding a few lines of code to your ETL script, as shown in the following examples. The code uses `enableUpdateCatalog` set to true, and also `updateBehavior` set to UPDATE_IN_DATABASE, which indicates to overwrite the schema and add new partitions in the Data Catalog during the job run.

Python

```
additionalOptions = {
    "enableUpdateCatalog": True,
    "updateBehavior": "UPDATE_IN_DATABASE"}
additionalOptions["partitionKeys"] = ["partition_key0", "partition_key1"]

sink = glueContext.write_dynamic_frame_from_catalog(frame=last_transform,
 database=<dst_db_name>,
    table_name=<dst_tbl_name>, transformation_ctx="write_sink",
    additional_options=additionalOptions)
job.commit()
```

Scala

```
val options = JsonOptions(Map(
    "path" -> outputPath,
    "partitionKeys" -> Seq("partition_0", "partition_1"),
    "enableUpdateCatalog" -> true))
val sink = glueContext.getCatalogSink(database = nameSpace, tableName = tableName,
 additionalOptions = options)
sink.writeDynamicFrame(df)
```

You can also set the `updateBehavior` value to `LOG` if you want to prevent your table schema from being overwritten, but still want to add the new partitions. The default value of `updateBehavior` is `UPDATE_IN_DATABASE`, so if you don't explicitly define it, then the table schema will be overwritten.

If `enableUpdateCatalog` is not set to true, regardless of whichever option selected for `updateBehavior`, the ETL job will not update the table in the Data Catalog.

Creating new tables

You can also use the same options to create a new table in the Data Catalog. You can specify the database and new table name using `setCatalogInfo`.

Python

```
sink = glueContext.getSink(connection_type="s3", path="s3://path/to/data",
    enableUpdateCatalog=True, updateBehavior="UPDATE_IN_DATABASE",
    partitionKeys=["partition_key0", "partition_key1"])
sink.setFormat("<format>")
sink.setCatalogInfo(catalogDatabase=<dst_db_name>, catalogTableName=<dst_tbl_name>)
sink.writeFrame(last_transform)
```

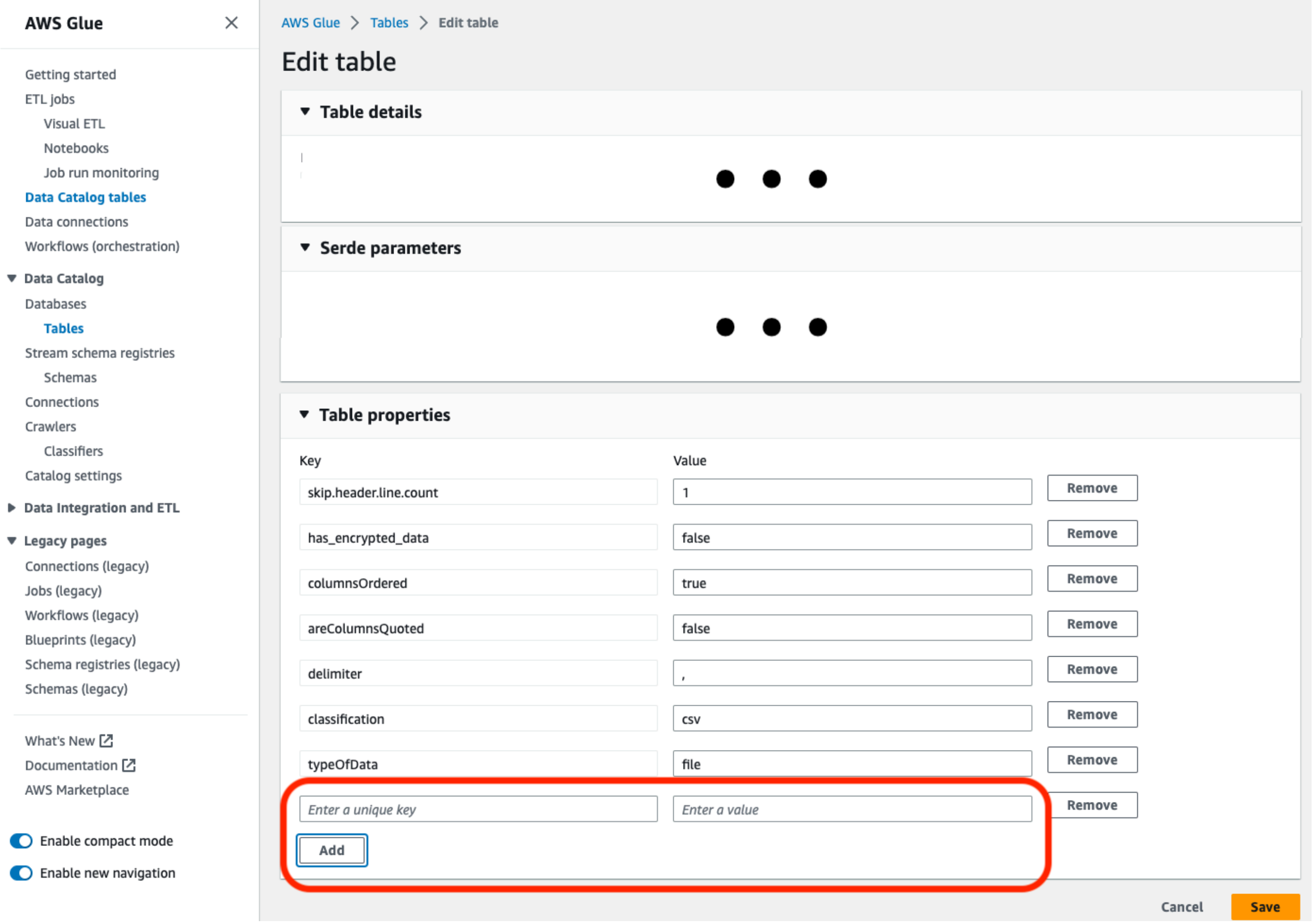
Scala

```
val options = JsonOptions(Map(
    "path" -> outputPath,
    "partitionKeys" -> Seq("<partition_1>", "<partition_2>"),
    "enableUpdateCatalog" -> true,
    "updateBehavior" -> "UPDATE_IN_DATABASE"))
val sink = glueContext.getSink(connectionType = "s3", connectionOptions =
 options).withFormat("<format>")
sink.setCatalogInfo(catalogDatabase = “<dst_db_name>”, catalogTableName =
 “<dst_tbl_name>”)
sink.writeDynamicFrame(df)
```

Restrictions

Take note of the following restrictions:

- Only Amazon Simple Storage Service (Amazon S3) targets are supported.
- The `enableUpdateCatalog` feature is not supported for governed tables.
- Only the following formats are supported: `json`, `csv`, `avro`, and `parquet`.
- To create or update tables with the `parquet` classification, you must utilize the AWS Glue optimized parquet writer for DynamicFrames. This can be achieved with one of the following:
  - If you're updating an existing table in the catalog with `parquet` classification, the table must have the `"useGlueParquetWriter"` table property set to `true` before you update it. You can set this property via the AWS Glue APIs/SDK, via the console or via an Athena DDL statement.


AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Legacy pages
Connections (legacy)
Jobs (legacy)
Workflows (legacy)
Blueprints (legacy)
Schema registries (legacy)
Schemas (legacy)
What's New
Documentation
AWS Marketplace
Enable compact mode
Enable new navigation
AWS Glue > Tables > Edit table
Edit table
Table details
Serde parameters
Table properties
Remove
Enter a unique key
Enter a value
Add
Cancel
Save

Key	Value
skip.header.line.count	1
has_encrypted_data	false
columnsOrdered	true
areColumnsQuoted	false
delimiter	,
classification	csv
typeOfData	file

Once the catalog table property is set, you can use the following snippet of code to update the catalog table with the new data:

```
glueContext.write_dynamic_frame.from_catalog(
    frame=frameToWrite,
    database="dbName",
    table_name="tableName",
    additional_options={
        "enableUpdateCatalog": True,
        "updateBehavior": "UPDATE_IN_DATABASE"
    }
)
```

- If the table doesn't already exist within catalog, you can utilize the `getSink()` method in your script with `connection_type="s3"` to add the table and its partitions to the catalog, along with writing the data to Amazon S3. Provide the appropriate `partitionKeys` and `compression` for your workflow.

```
s3sink = glueContext.getSink(
    path="s3://bucket/folder/",
    connection_type="s3",
    updateBehavior="UPDATE_IN_DATABASE",
    partitionKeys=[],
    compression="snappy",
    enableUpdateCatalog=True
)

s3sink.setCatalogInfo(
    catalogDatabase="dbName", catalogTableName="tableName"
)

s3sink.setFormat("parquet", useGlueParquetWriter=True)
s3sink.writeFrame(frameToWrite)
```

  - The `glueparquet` format value is a legacy method of enabling the AWS Glue parquet writer.
- When the `updateBehavior` is set to `LOG`, new partitions will be added only if the `DynamicFrame` schema is equivalent to or contains a subset of the columns defined in the Data Catalog table's schema.
- Schema updates are not supported for non-partitioned tables (not using the "partitionKeys" option).
- Your partitionKeys must be equivalent, and in the same order, between your parameter passed in your ETL script and the partitionKeys in your Data Catalog table schema.
- This feature currently does not yet support updating/creating tables in which the updating schemas are nested (for example, arrays inside of structs).

For more information, see the section called “AWS Glue for Spark”.

Working with MongoDB connections in ETL jobs

You can create a connection for MongoDB and then use that connection in your AWS Glue job. For more information, see the section called “MongoDB connections” in the AWS Glue programming guide. The connection `url`, `username` and `password` are stored in the MongoDB connection. Other options can be specified in your ETL job script using the `additionalOptions` parameter of `glueContext.getCatalogSource`. The other options can include:

- `database`: (Required) The MongoDB database to read from.

- `collection`: (Required) The MongoDB collection to read from.

By placing the `database` and `collection` information inside the ETL job script, you can use the same connection for in multiple jobs.

1. Create an AWS Glue Data Catalog connection for the MongoDB data source. See "connectionType": "mongodb" for a description of the connection parameters. You can create the connection using the console, APIs or CLI.
2. Create a database in the AWS Glue Data Catalog to store the table definitions for your MongoDB data. See Creating databases for more information.
3. Create a crawler that crawls the data in the MongoDB using the information in the connection to connect to the MongoDB. The crawler creates the tables in the AWS Glue Data Catalog that describe the tables in the MongoDB database that you use in your job. See Using crawlers to populate the Data Catalog for more information.
4. Create a job with a custom script. You can create the job using the console, APIs or CLI. For more information, see Adding Jobs in AWS Glue.
5. Choose the data targets for your job. The tables that represent the data target can be defined in your Data Catalog, or your job can create the target tables when it runs. You choose a target location when you author the job. If the target requires a connection, the connection is also referenced in your job. If your job requires multiple data targets, you can add them later by editing the script.
6. Customize the job-processing environment by providing arguments for your job and generated script.

   Here is an example of creating a `DynamicFrame` from the MongoDB database based on the table structure defined in the Data Catalog. The code uses `additionalOptions` to provide the additional data source information:

   Scala

   ```
   val resultFrame: DynamicFrame = glueContext.getCatalogSource(
           database = catalogDB,
           tableName = catalogTable,
           additionalOptions = JsonOptions(Map("database" -> DATABASE_NAME,
                   "collection" -> COLLECTION_NAME))
         ).getDynamicFrame()
   ```

Python

```
glue_context.create_dynamic_frame_from_catalog(
        database = catalogDB,
        table_name = catalogTable,
        additional_options = {"database":"database_name",
            "collection":"collection_name"})
```

7. Run the job, either on-demand or through a trigger.

Optimizing query performance using column statistics

You can compute column-level statistics for AWS Glue Data Catalog tables in data formats such as Parquet, ORC, JSON, ION, CSV, and XML without setting up additional data pipelines. Column statistics help you to understand data profiles by getting insights about values within a column.

Data Catalog supports generating statistics for column values such as minimum value, maximum value, total null values, total distinct values, average length of values, and total occurrences of true values. AWS analytical services such as Amazon Redshift and Amazon Athena can use these column statistics to generate query execution plans, and choose the optimal plan that improves query performance.

There are three scenarios for generating column statistics:

Auto

AWS Glue supports automatic column statistics generation at the catalog-level so that it can automatically generate statistics for new tables in the AWS Glue Data Catalog.

Scheduled

AWS Glue supports scheduling column statistics generation so that it can be run automatically on a recurring schedule.

With scheduled statistics computation, the column statistics task updates the overall table-level statistics, such as min, max, and avg with the new statistics, providing query engines with accurate and up-to-date statistics to optimize query execution.

On-demand

Use this option to generate column statistics on-demand whenever needed. This is useful for ad-hoc analysis or when statistics need to be computed immediately.

You can configure to run column statistics generation task using AWS Glue console, AWS CLI, and AWS Glue API operations. When you initiate the process, AWS Glue starts a Spark job in the background and updates the AWS Glue table metadata in the Data Catalog. You can view column statistics using AWS Glue console or AWS CLI or by calling the GetColumnStatisticsForTable API operation.

> **Note**
>
> If you're using Lake Formation permissions to control access to the table, the role assumed by the column statistics task requires full table access to generate statistics.

Topics

- Prerequisites for generating column statistics
- Automatic column statistics generation
- Generating column statistics on a schedule
- Generating column statistics on demand
- Viewing column statistics
- Viewing column statistics task runs
- Stopping column statistics task run
- Deleting column statistics
- Considerations and limitations

Prerequisites for generating column statistics

To generate or update column statistics, the statistics generation task assumes an AWS Identity and Access Management (IAM) role on your behalf. Based on the permissions granted to the role, the column statistics generation task can read the data from the Amazon S3 data store.

When you configure the column statistics generation task, AWS Glue allows you to create a role that includes the `AWSGlueServiceRole` AWS managed policy plus the required inline policy for the specified data source.

If you specify an existing role for generating column statistics, ensure that it includes the `AWSGlueServiceRole` policy or equivalent (or a scoped down version of this policy), plus the required inline policies. Follow these steps to create a new IAM role:

Note

To generate statistics for tables managed by Lake Formation, the IAM role used to generate statistics requires full table access.

When you configure the column statistics generation task, AWS Glue allows you to create a role that includes the `AWSGlueServiceRole` AWS managed policy plus the required inline policy for the specified data source. You can also create a role and attach the the permissions listed in the policy below, and add that role to the column statistics generation task.

To create an IAM role for generating column statistics

1. To create an IAM role, see Create an IAM role for AWS Glue.
2. To update an existing role, in the IAM console, go to the IAM role that is being used by the generate column statistics process.
3. In the **Add permissions** section, choose **Attach policies**. In the newly opened browser window, choose `AWSGlueServiceRole` AWS managed policy.
4. You also need to include permissions to read data from the Amazon S3 data location.

   In the **Add permissions** section, choose **Create policy**. In the newly opened browser window, create a new policy to use with your role.
5. In the **Create policy** page, choose the **JSON** tab. Copy the following `JSON` code into the policy editor field.

   Note

   In the following policies, replace account ID with a valid AWS account, and replace `region` with the Region of the table, and `bucket-name` with the Amazon S3 bucket name.

   JSON

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "S3BucketAccess",
            "Effect": "Allow",
            "Action": [
                "s3:ListBucket",
                "s3:GetObject"
            ],
            "Resource": [
             "arn:aws:s3:::amzn-s3-demo-bucket/*",
      "arn:aws:s3:::amzn-s3-demo-bucket"
            ]
        }
     ]
}
```

6. (Optional) If you're using Lake Formation permissions to provide access to your data, the IAM role requires `lakeformation:GetDataAccess` permissions.

   JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "LakeFormationDataAccess",
      "Effect": "Allow",
      "Action": "lakeformation:GetDataAccess",
      "Resource": [
        "*"
      ]
    }
  ]
}
```

   If the Amazon S3 data location is registered with Lake Formation, and the IAM role assumed by the column statistics generation task doesn't have `IAM_ALLOWED_PRINCIPALS` group permissions granted on the table, the role requires Lake Formation `ALTER` and `DESCRIBE` permissions on the table. The role used for registering the Amazon S3 bucket requires Lake Formation `INSERT` and `DELETE` permissions on the table.

   If the Amazon S3 data location is not registered with Lake Formation, and the IAM role doesn't have `IAM_ALLOWED_PRINCIPALS` group permissions granted on the table, the role requires Lake Formation `ALTER`, `DESCRIBE`, `INSERT` and `DELETE` permissions on the table.

7. If you've enabled the catalog-level `Automatic statistics generation` option, the IAM role must have the `glue:UpdateCatalog` permission or the Lake Formation `ALTER CATALOG` permission on the default Data Catalog. You can use the `GetCatalog` operation to verify the catalog properties.
8. (Optional) The column statistics generation task that writes encrypted Amazon CloudWatch Logs requires the following permissions in the key policy.

   JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "CWLogsKmsPermissions",
      "Effect": "Allow",
      "Action": [
        "logs:CreateLogGroup",
        "logs:CreateLogStream",
        "logs:PutLogEvents",
        "logs:AssociateKmsKey"
      ],
      "Resource": [
        "arn:aws:logs:us-east-1:111122223333:log-group:/aws-glue:*"
      ]
    },
    {
      "Sid": "KmsPermissions",
      "Effect": "Allow",
      "Action": [
        "kms:GenerateDataKey",
        "kms:Decrypt",
        "kms:Encrypt"
      ],
      "Resource": [
        "arn:aws:kms:us-east-1:111122223333:key/arn of key used for ETL
 cloudwatch encryption"
      ],
        "Condition": {
          "StringEquals": {
            "kms:ViaService": [
              "glue.us-east-1.amazonaws.com"
            ]
          }
        }
      }
    ]
  }
```

9. The role you use to run column statistics must have the `iam:PassRole` permission on the role.

   JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "iam:PassRole"
      ],
      "Resource": [
        "arn:aws:iam::111122223333:role/columnstats-role-name"
      ]
    }
  ]
}
```

10. When you create an IAM role for generating column statistics, that role must also have the following trust policy that enables the service to assume the role.

    JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "TrustPolicy",
      "Effect": "Allow",
      "Principal": {
        "Service": "glue.amazonaws.com"
      },
      "Action": "sts:AssumeRole"
    }
  ]
}
```

Automatic column statistics generation

Automatic generation of column statistics allows you to schedule and automatically compute statistics on new tables in the AWS Glue Data Catalog. When you enable automatic statistics generation, the Data Catalog discovers new tables with specific data formats such as Parquet, JSON, CSV, XML, ORC, ION, and Apache Iceberg, along with their individual bucket paths. With a one-time catalog configuration, the Data Catalog generates statistics for these tables.

Data lake administrators can configure the statistics generation by selecting the default catalog in the Lake Formation console, and enabling table statistics using the `Optimization configuration` option. When you create new tables or update existing tables in the Data Catalog, the Data Catalog collects the number of distinct values (NDVs) for Apache Iceberg tables, and additional statistics such as the number of nulls, maximum, minimum, and average length for other supported file formats on a weekly basis.

If you have configured statistics generation at the table-level or if you have previously deleted the statistics generation settings for a table, those table-specific settings take precedence over the default catalog settings for automatic column statistics generation.

Automatic statistics generation task analyzes 50% of records in the tables to calculate statistics. Automatic column statistics generation ensures that the Data Catalog maintains weekly metrics that can be used by query engines like Amazon Athena and Amazon Redshift Spectrum for improved query performance and potential cost savings. It allows scheduling statistics generation using AWS Glue APIs or the console, providing an automated process without manual intervention.

Topics

- Enabling catalog-level automatic statistics generation
- Viewing automated table-level settings
- Disabling catalog-level column statistics generation

Enabling catalog-level automatic statistics generation

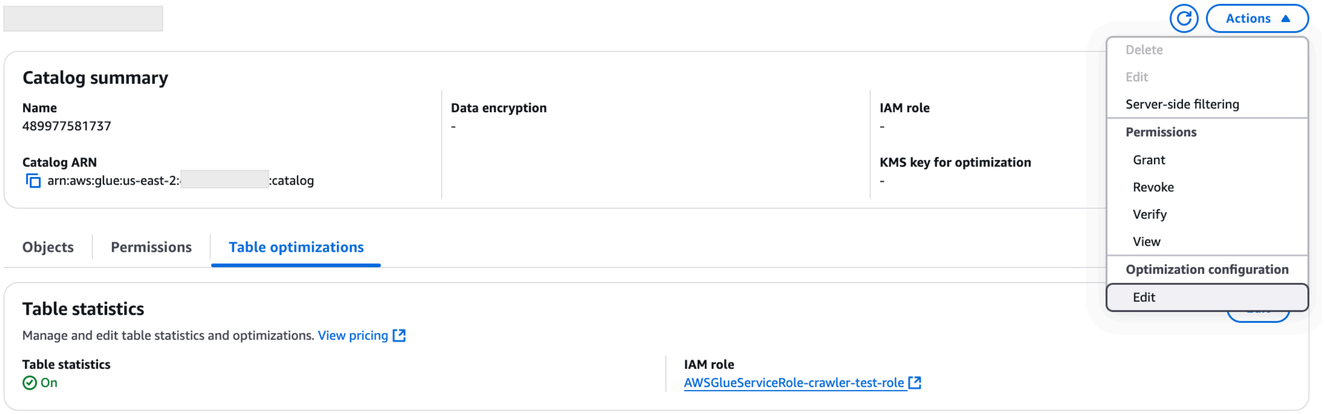
You can enable the automatic column statistics generation for all new Apache Iceberg tables and tables in non-OTF table (Parquet, JSON, CSV, XML, ORC, ION) formats in the Data Catalog. After creating the table, you can also explicitly update the column statistics settings manually.

To update the Data Catalog settings to enable catalog-level, the IAM role used must have the `glue:UpdateCatalog` permission or AWS Lake Formation `ALTER CATALOG` permission on the root catalog. You can use `GetCatalog` API to verify the catalog properties.

AWS Management Console

To enable the automatic column statistics generation at the account-level

1. Open the Lake Formation console at https://console.aws.amazon.com/lakeformation/.
2. On the left navigation bar, choose **Catalogs**.
3. On the **Catalog summary** page, choose **Edit** under **Optimization configuration**.



[redacted]

Catalog summary

Name: 489977581737

Data encryption: -

IAM role: -

Catalog ARN: arn:aws:glue:us-east-2:[redacted]:catalog

KMS key for optimization: -

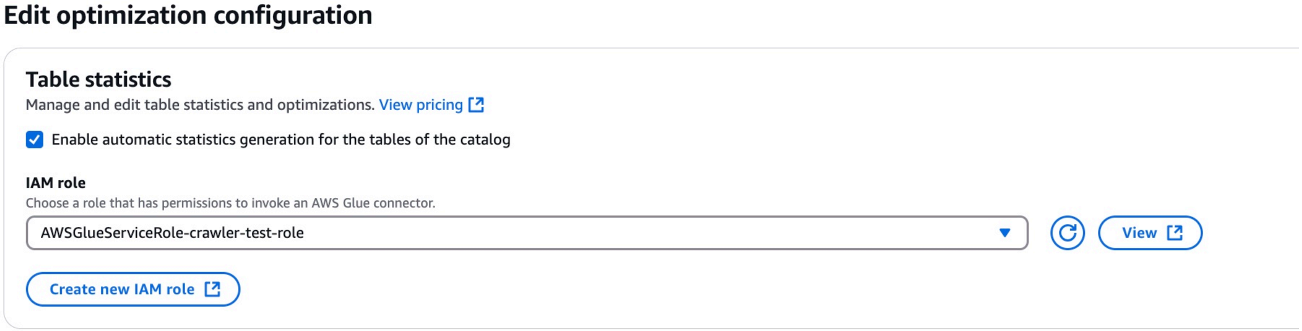
Table statistics

Manage and edit table statistics and optimizations. View pricing

Table statistics: On

IAM role: AWSGlueServiceRole-crawler-test-role

4. On the **Table optimization configuration** page, choose the **Enable automatic statistics generation for the tables of the catalog** option.



Edit optimization configuration

Table statistics

Manage and edit table statistics and optimizations. View pricing

☑ Enable automatic statistics generation for the tables of the catalog

IAM role

Choose a role that has permissions to invoke an AWS Glue connector.

AWSGlueServiceRole-crawler-test-role

5. Choose an existing IAM role or create a new one that has the necessary permissions to run the column statistics task.
6. Choose **Submit**.

AWS CLI

You can also enable catalog-level statistics collection through the AWS CLI. To configure table-level statistics collection using AWS CLI, run the following command:

```
aws glue update-catalog --cli-input-json '{
    "name": "123456789012",
    "catalogInput": {
        "description": "Updating root catalog with role arn",
        "catalogProperties": {
            "customProperties": {
                "ColumnStatistics.RoleArn": "arn:aws:iam::"123456789012":role/
service-role/AWSGlueServiceRole",
                "ColumnStatistics.Enabled": "true"
            }
        }
    }
}'
```

The above command calls AWS Glue's `UpdateCatalog` operation, which takes in a `CatalogProperties` structure with the following key-value pairs for catalog-level statistics generation:

- ColumnStatistics.RoleArn – IAM role ARN to be used for all tasks triggered for Catalog-level statistics generation
- ColumnStatistics.Enabled – Boolean indicating whether the catalog-level settings is enabled or disabled

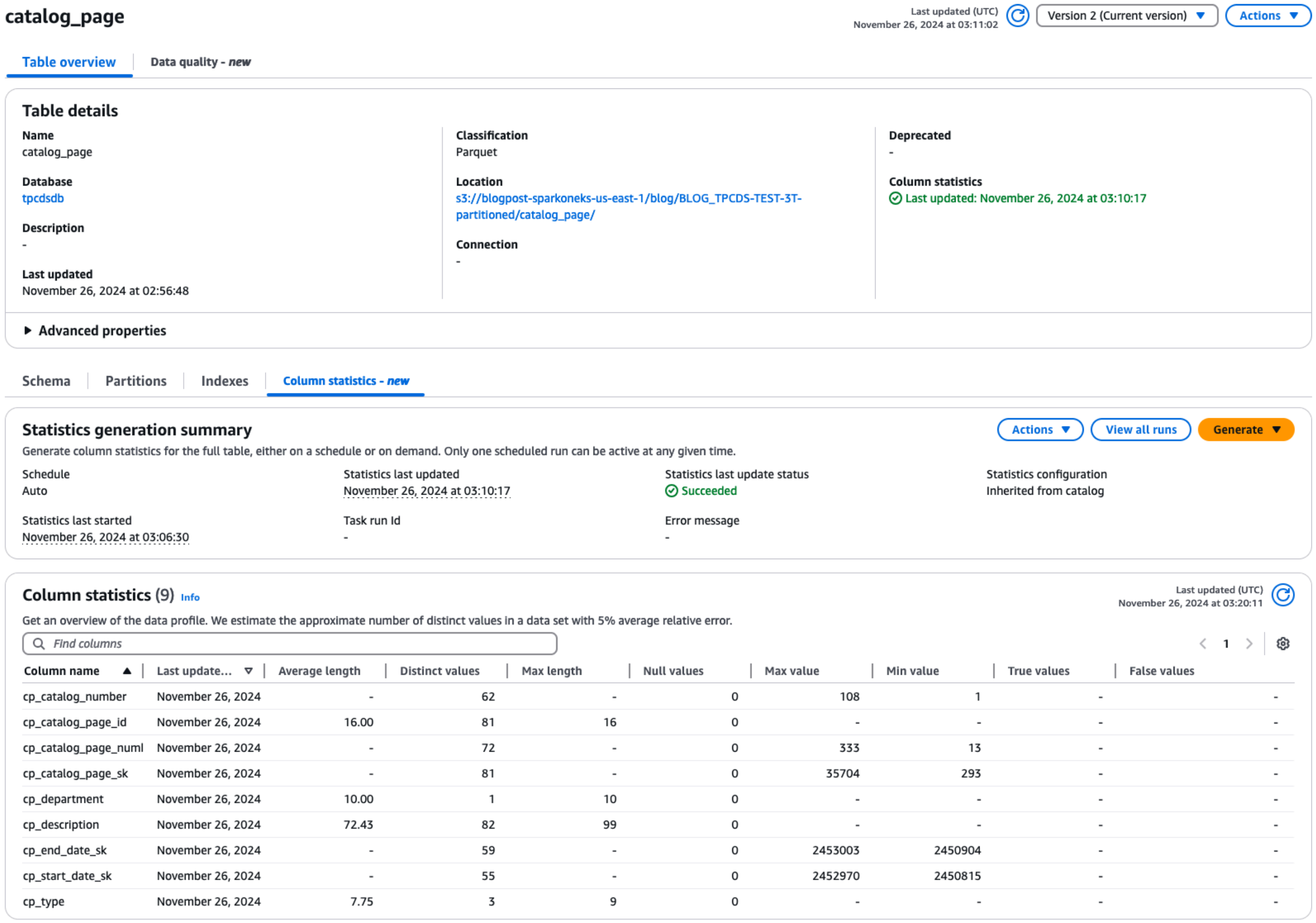
Viewing automated table-level settings

When catalog-level statistics collection is enabled, anytime an Apache Hive table or Apache Iceberg table is created or updated via the `CreateTable` or `UpdateTable` APIs through AWS Management Console, SDK, or AWS Glue crawler, an equivalent table level setting is created for that table.

Tables with automatic statistics generation enabled must follow one of following properties:

- Use an `InputSerdeLibrary` that begins with org.apache.hadoop and `TableType` equals `EXTERNAL_TABLE`
- Use an `InputSerdeLibrary` that begins with `com.amazon.ion` and `TableType` equals `EXTERNAL_TABLE`
- Contain table_type: "ICEBERG" in it’s parameters structure.

After you create or update a table, you can verify the table details to confirm the statistics generation. The `Statistics generation summary` shows the `Schedule` property set as `AUTO` and `Statistics configuration` value is `Inherited from catalog`. Any table setting with the following setting would be automatically triggered by Glue internally.



catalog_page

Last updated (UTC)
November 26, 2024 at 03:11:02

Table details

Name
catalog_page

Database
tpcdsdb

Description
-

Last updated
November 26, 2024 at 02:56:48

Classification
Parquet

Location
s3://blogpost-sparkoneks-us-east-1/blog/BLOG_TPCDS-TEST-3T-partitioned/catalog_page/

Connection
-

Deprecated
-

Column statistics
Last updated: November 26, 2024 at 03:10:17

Statistics generation summary

Generate column statistics for the full table, either on a schedule or on demand. Only one scheduled run can be active at any given time.

Schedule
Auto

Statistics last updated
November 26, 2024 at 03:10:17

Statistics last update status
Succeeded

Statistics configuration
Inherited from catalog

Statistics last started
November 26, 2024 at 03:06:30

Task run Id
-

Error message
-

Column statistics (9)

Last updated (UTC)
November 26, 2024 at 03:20:11

Get an overview of the data profile. We estimate the approximate number of distinct values in a data set with 5% average relative error.

Column name	Last update...	Average length	Distinct values	Max length	Null values	Max value	Min value	True values	False values
cp_catalog_number	November 26, 2024	-	62	-	0	108	1	-	-
cp_catalog_page_id	November 26, 2024	16.00	81	16	0	-	-	-	-
cp_catalog_page_numl	November 26, 2024	-	72	-	0	333	13	-	-
cp_catalog_page_sk	November 26, 2024	-	81	-	0	35704	293	-	-
cp_department	November 26, 2024	10.00	1	10	0	-	-	-	-
cp_description	November 26, 2024	72.43	82	99	0	-	-	-	-
cp_end_date_sk	November 26, 2024	-	59	-	0	2453003	2450904	-	-
cp_start_date_sk	November 26, 2024	-	55	-	0	2452970	2450815	-	-
cp_type	November 26, 2024	7.75	3	9	0	-	-	-	-

Disabling catalog-level column statistics generation

You can disable automatic column statistics generation for new tables using the AWS Lake Formation console, the `glue:UpdateCatalogSettings` API, or the `glue:DeleteColumnStatisticsTaskSettings` API.

To disable the automatic column statistics generation at the account-level

1. Open the Lake Formation console at https://console.aws.amazon.com/lakeformation/.
2. On the left navigation bar, choose **Catalogs**.
3. On the **Catalog summary** page, choose **Edit** under **Optimization configuration**.
4. On the **Table optimization configuration** page, unselect the **Enable automatic statistics generation for the tables of the catalog** option.
5. Choose **Submit**.

Generating column statistics on a schedule

Follow these steps to configure a schedule for generating column statistics in the AWS Glue Data Catalog using the AWS Glue console, the AWS CLI, or the CreateColumnStatisticsTaskSettings operation.

Console

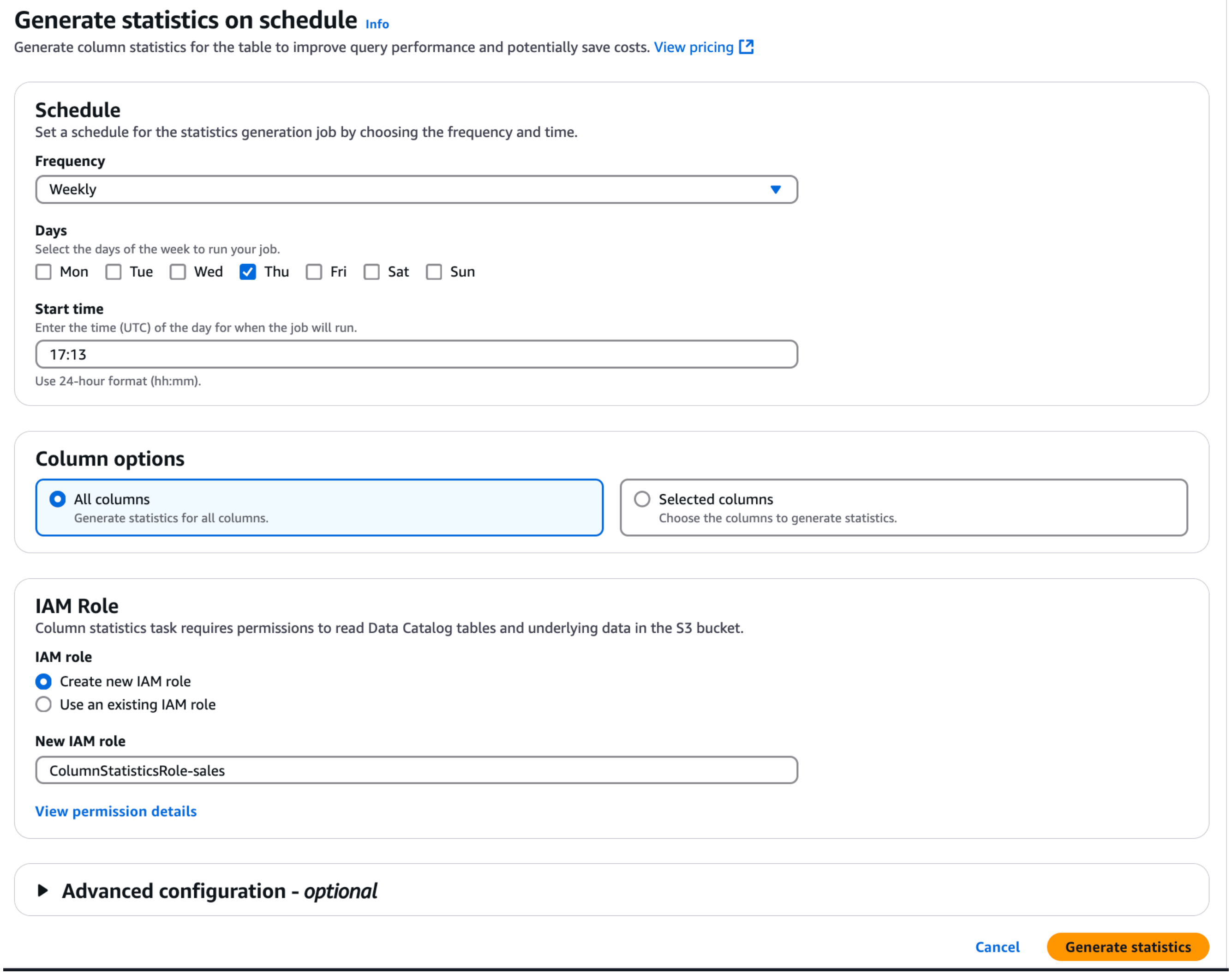
To generate column statistics using the console

1. Sign in to the AWS Glue console at https://console.aws.amazon.com/glue/.
2. Choose Data Catalog tables.
3. Choose a table from the list.
4. Choose **Column statistics** tab in the lower section of the **Tables** page.
5. You can also choose **Generate on schedule** under **Column statistics** from **Actions**.
6. On the **Generate statistics on schedule** page, configure a recurring schedule for running the column statistics task by choosing the frequency and start time. You can choose the frequency to be hourly, daily, weekly, or define a cron expression to specify the schedule.

   A cron expression is a string representing a schedule pattern, consisting of 6 fields separated by spaces: * * * * * <minute> <hour> <day of month> <month> <day of week>

<year> For example, to run a task every day at midnight, the cron expression would be: 0 0 * * ? *

For more information, see Cron expressions.



Generate statistics on schedule Info

Generate column statistics for the table to improve query performance and potentially save costs. View pricing

Schedule

Set a schedule for the statistics generation job by choosing the frequency and time.

Frequency: Weekly

Days

Select the days of the week to run your job.

☐ Mon ☐ Tue ☐ Wed ☑ Thu ☐ Fri ☐ Sat ☐ Sun

Start time

Enter the time (UTC) of the day for when the job will run.

17:13

Use 24-hour format (hh:mm).

Column options

● All columns – Generate statistics for all columns.

○ Selected columns – Choose the columns to generate statistics.

IAM Role

Column statistics task requires permissions to read Data Catalog tables and underlying data in the S3 bucket.

IAM role

● Create new IAM role

○ Use an existing IAM role

New IAM role: ColumnStatisticsRole-sales

7. Next, choose the column option to generate statistics.

   - **All columns** – Choose this option to generate statistics for all columns in the table.
   - **Selected columns** – Choose this option to generate statistics for specific columns. You can select the columns from the drop-down list.

8. Choose an IAM role or create an existing role that has permissions to generate statistics. AWS Glue assumes this role to generate column statistics.

   A quicker approach is to let the AWS Glue console to create a role for you. The role that it creates is specifically for generating column statistics, and includes the

`AWSGlueServiceRole` AWS managed policy plus the required inline policy for the specified data source.

If you specify an existing role for generating column statistics, ensure that it includes the `AWSGlueServiceRole` policy or equivalent (or a scoped down version of this policy), plus the required inline policies.

9. (Optional) Next, choose a security configuration to enable at-rest encryption for logs.
10. (Optional) You can choose a sample size by indicating only a specific percent of rows from the table to generate statistics. The default is all rows. Use the up and down arrows to increase or decrease the percent value.

    We recommend to include all rows in the table to compute accurate statistics. Use sample rows to generate column statistics only when approximate values are acceptable.
11. Choose **Generate statistics** to run the column statistics generation task.

AWS CLI

You can use the following AWS CLI example to create a column statistics generation schedule. The database-name, table-name, and role are required parameters, and optional parameters are schedule, column-name-list, catalog-id, sample-size, and security-configuration.

```
aws glue create-column-statistics-task-settings \
 --database-name 'database_name' \
 --table-name table_name \
 --role 'arn:aws:iam::123456789012:role/stats-role' \
 --schedule 'cron(0 0-5 14 * * ?)' \
 --column-name-list 'col-1' \
 --catalog-id '123456789012' \
 --sample-size '10.0 ' \
 --security-configuration 'test-security'
```

You can generate column statistics also by calling the StartColumnStatisticsTaskRun operation.

Managing the schedule for column statistics generation

You can manage the scheduling operations such as updating, starting, stopping, and deleting schedules for the column statistics generation in AWS Glue. You can use AWS Glue console, AWS CLI, or AWS Glue column statistics API operations to perform these tasks.

Topics

- Updating the column statistics generation schedule
- Stopping the schedule for column statistics generation
- Resuming the schedule for column statistics generation
- Deleting column statistics generation schedule

Updating the column statistics generation schedule

You can update the schedule to trigger the column statistics generation task after it has been created. You can use the AWS Glue console, AWS CLI, or run the UpdateColumnStatisticsTaskSettings operation to update the schedule for a table. You can modify the parameters of an existing schedule, such as the schedule type (on-demand, or scheduled) and other optional parameters.

AWS Management Console

To update the settings for a column statistics generation task

1. Sign in to the AWS Glue console at https://console.aws.amazon.com/glue/.
2. Choose the table that you want to update from the tables list.
3. In the lower section of the table details page, choose **Column statistics**.
4. Under **Actions**, choose **Edit** to update the schedule.
5. Make the desired changes to the schedule, and choose **Save**.

AWS CLI

If you are not using AWS Glue's statistics generation feature in the console, you can manually update the schedule using the `update-column-statistics-task-settings` command. The following example shows how to update column statistics using AWS CLI.

```
aws glue update-column-statistics-task-settings \
 --database-name 'database_name' \
 --table-name 'table_name' \
 --role arn:aws:iam::123456789012:role/stats_role \
 --schedule 'cron(0 0-5 16 * * ?)' \
 --column-name-list 'col-1' \
 --sample-size '20.0' \
 --catalog-id '123456789012'\
 --security-configuration 'test-security'
```

Stopping the schedule for column statistics generation

If you no longer need the incremental statistics, you can stop the scheduled generation to save resources and costs. Pausing the schedule doesn't impact the previously generated statistics. You can resume the schedule at your convenience.

AWS Management Console

To stop the schedule for a column statistics generation task

1. On AWS Glue console, choose **Tables** under Data Catalog.
2. Select a table with column statistics.
3. On the **Table details** page, choose **Column statistics**.
4. Under **Actions**, choose **Scheduled generation**, **Pause**.
5. Choose **Pause** to confirm.

AWS CLI

To stop a column statistics task run schedule using the AWS CLI, you can use the following command:

```
aws glue stop-column-statistics-task-run-schedule \
 --database-name ''database_name' \
 --table-name 'table_name'
```

Replace the `database_name` and the `table_name` with the actual names of the database and table for which you want to stop the column statistics task run schedule.

Resuming the schedule for column statistics generation

If you've paused the statistics generation schedule, AWS Glue allows you to resume the schedule at your convenience. You can resume the schedule using the AWS Glue console, AWS CLI, or the StartColumnStatisticsTaskRunSchedule operation.

AWS Management Console

To resume the schedule for column statistics generation

1. On AWS Glue console, choose **Tables** under Data Catalog.
2. Select a table with column statistics.
3. On the **Table details** page, choose **Column statistics**.
4. Under **Actions**, choose **Scheduled generation**, and choose **Resume**.
5. Choose **Resume**to confirm.

AWS CLI

Replace the `database_name` and the `table_name` with the actual names of the database and table for which you want to stop the column statistics task run schedule.

```
aws glue start-column-statistics-task-run-schedule \
 --database-name 'database_name' \
 --table-name 'table_name'
```

Deleting column statistics generation schedule

While maintaining up-to-date statistics is generally recommended for optimal query performance, there are specific use cases where removing the automatic generation schedule might be beneficial.

- If the data remains relatively static, the existing column statistics may remain accurate for an extended period, reducing the need for frequent updates. Deleting the schedule can prevent unnecessary resource consumption and overhead associated with regenerating statistics on unchanging data.
- When manual control over statistics generation is preferred. By deleting the automatic schedule, administrators can selectively update column statistics at specific intervals or after significant data changes, aligning the process with their maintenance strategies and resource allocation needs.

AWS Management Console

To delete the schedule for column statistics generation

1. On AWS Glue console, choose **Tables** under Data Catalog.
2. Select a table with column statistics.
3. On the **Table details** page, choose **Column statistics**.
4. Under **Actions**, choose **Scheduled generation**, **Delete**.
5. Choose **Delete**to confirm.

AWS CLI

Replace the `database_name` and the `table_name` with the actual names of the database and table for which you want to stop the column statistics task run schedule.

You can delete column statistics schedule using the DeleteColumnStatisticsTaskSettings API operation or AWS CLI. The following example shows how to delete the schedule for generating column statistics using AWS Command Line Interface (AWS CLI).

```
aws glue delete-column-statistics-task-settings \
    --database-name 'database_name' \
    --table-name 'table_name'
```

Generating column statistics on demand

You can run the column statistics task for the AWS Glue Data Catalog tables task on-demand without a set schedule. This option is useful for ad-hoc analysis or when statistics need to be computed immediately.

Follow these steps to generate column statistics on demand for the Data Catalog tables using AWS Glue console or AWS CLI.

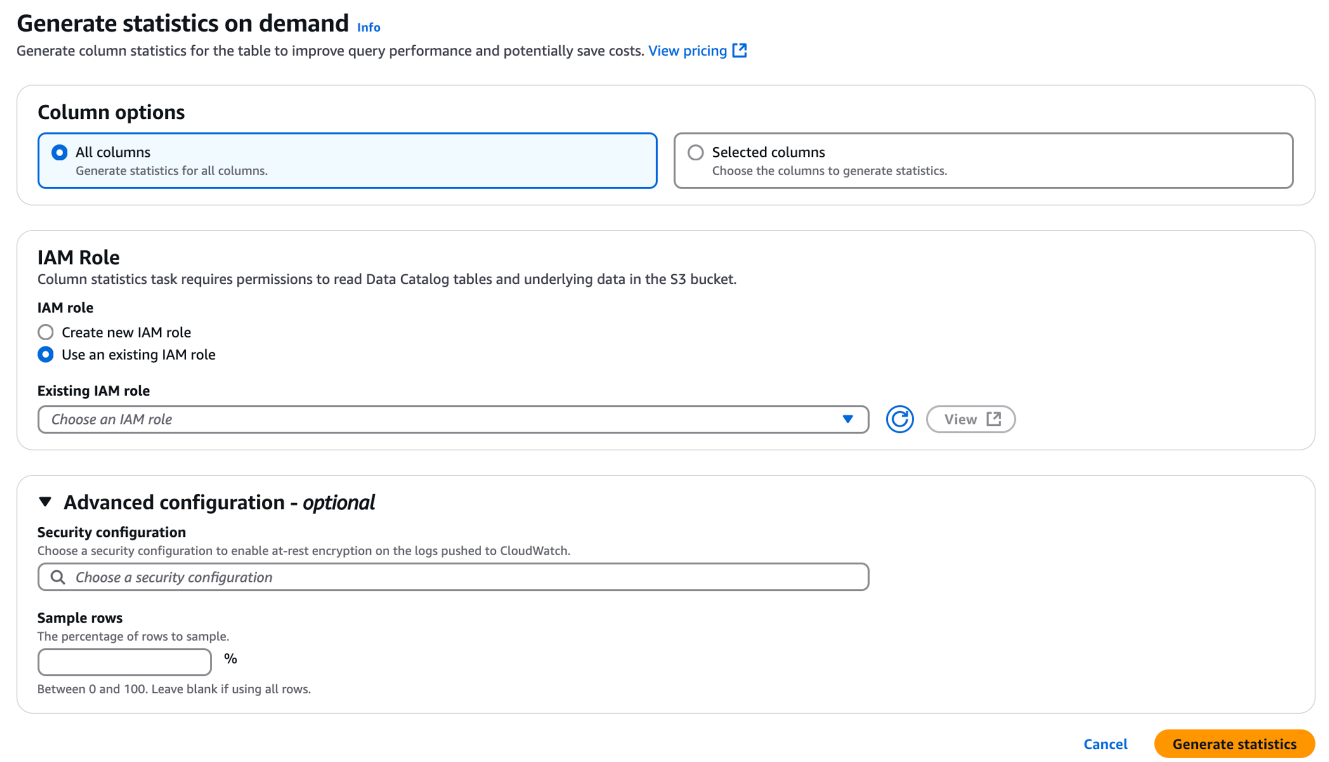
AWS Management Console

To generate column statistics using the console

1. Sign in to the AWS Glue console at https://console.aws.amazon.com/glue/.
2. Choose Data Catalog tables.

3. Choose a table from the list.
4. Choose **Generate statistics** under **Actions** menu.

   You can also choose **Generate**, **Generate on demand** option under **Column statistics** tab in the lower section of the **Table** page.
5. Follow steps 7 - 11 in the [Generating column statistics on a schedule]() to generate column statistics for the table.
6. On the **Generate statistics** page, specify the following options:



Generate statistics on demand Info

Generate column statistics for the table to improve query performance and potentially save costs. View pricing

Column options

● All columns
Generate statistics for all columns.

○ Selected columns
Choose the columns to generate statistics.

IAM Role

Column statistics task requires permissions to read Data Catalog tables and underlying data in the S3 bucket.

IAM role

○ Create new IAM role
● Use an existing IAM role

Existing IAM role

Choose an IAM role

Advanced configuration - *optional*

Security configuration

Choose a security configuration to enable at-rest encryption on the logs pushed to CloudWatch.

Choose a security configuration

Sample rows

The percentage of rows to sample.

%

Between 0 and 100. Leave blank if using all rows.

- **All columns** – Choose this option to generate statistics for all columns in the table.
- **Selected columns** – Choose this option to generate statistics for specific columns. You can select the columns from the drop-down list.
- **IAM role** –Choose **Create a new IAM role** that has the required permission policies to run the column statistics generation task. Choose View permission details to review the policy statement. You can also select an IAM role from the list. For more information about the required permissions, see [Prerequisites for generating column statistics]().

  AWS Glue assumes the permissions of the role that you specify to generate statistics.

For more information about providing roles for AWS Glue, see Identity-based policies for AWS Glue..

- (Optional) Next, choose a security configuration to enable at-rest encryption for logs.
- **Sample rows** – Choose only a specific percent of rows from the table to generate statistics. The default is all rows. Use the up and down arrows to increase or decrease the percent value.

> **Note**
>
> We recommend to include all rows in the table to compute accurate statistics. Use sample rows to generate column statistics only when approximate values are acceptable.

Choose **Generate statistics** to run the task.

AWS CLI

This command will trigger an column statistics task run for the specified table. You need to provide the database name, table name, an IAM role with permissions to generate statistics, and optionally provide column names and a sample size percentage for the statistics computation.

```
aws glue start-column-statistics-task-run \
    --database-name 'database_name \
    --table-name 'table_name' \
    --role 'arn:aws:iam::123456789012:role/stats-role' \
    --column-name 'col1','col2'  \
    --sample-size 10.0
```

This command will start a task to generate column statistics for the specified table.

Updating column statistics on demand

Maintaining up-to-date column statistics is crucial for the query optimizer to generate efficient execution plans, ensuring improved query performance, reduced resource consumption, and better overall system performance. This process is particularly important after significant data changes, such as bulk loads or extensive modifications, which can render existing statistics obsolete.

You need to explicitly run the **Generate statistics** task from the AWS Glue console to refresh the column statistics. Data Catalog doesn't automatically refresh the statistics.

If you are not using AWS Glue's statistics generation feature in the console, you can manually update column statistics using the UpdateColumnStatisticsForTable API operation or AWS CLI. The following example shows how to update column statistics using AWS CLI.

```
aws glue update-column-statistics-for-table --cli-input-json:

{
    "CatalogId": "111122223333",
    "DatabaseName": "database_name",
    "TableName": "table_name",
    "ColumnStatisticsList": [
        {
            "ColumnName": "col1",
            "ColumnType": "Boolean",
            "AnalyzedTime": "1970-01-01T00:00:00",
            "StatisticsData": {
                "Type": "BOOLEAN",
                "BooleanColumnStatisticsData": {
                    "NumberOfTrues": 5,
                    "NumberOfFalses": 5,
                    "NumberOfNulls": 0
                }
            }
        }
    ]
}
```

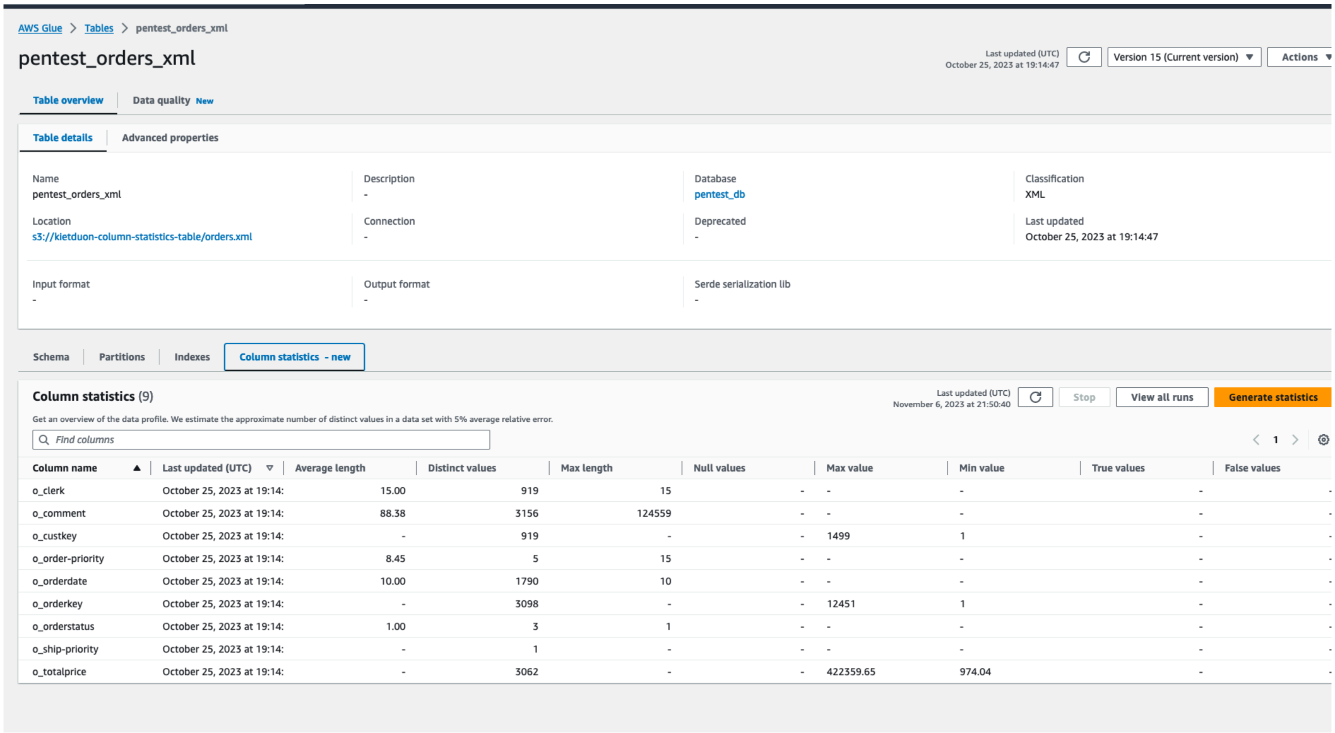
Viewing column statistics

After generating the statistics successfully, Data Catalog stores this information for the cost-based optimizers in Amazon Athena and Amazon Redshift to make optimal choices when running queries. The statistics varies based on the type of the column.

AWS Management Console

To view column statistics for a table

- After running column statistics task, the **Column statistics** tab on the **Table details** page shows the statistics for the table.



AWS Glue > Tables > pentest_orders_xml
Version 15 (Current version)
Actions
Table overview
Data quality New
Table details
Advanced properties
Schema
Partitions
Indexes
Column statistics - new
Stop
View all runs
Generate statistics
Find columns
1

pentest_orders_xml

Last updated (UTC) October 25, 2023 at 19:14:47

Name: pentest_orders_xml

Description: -

Database: pentest_db

Classification: XML

Location: s3://kietduon-column-statistics-table/orders.xml

Connection: -

Deprecated: -

Last updated: October 25, 2023 at 19:14:47

Input format: -

Output format: -

Serde serialization lib: -

Column statistics (9)

Last updated (UTC) November 6, 2023 at 21:50:40

Get an overview of the data profile. We estimate the approximate number of distinct values in a data set with 5% average relative error.

Column name	Last updated (UTC)	Average length	Distinct values	Max length	Null values	Max value	Min value	True values	False values
o_clerk	October 25, 2023 at 19:14:	15.00	919	15	-	-	-	-	-
o_comment	October 25, 2023 at 19:14:	88.38	3156	124559	-	-	-	-	-
o_custkey	October 25, 2023 at 19:14:	-	919	-	-	1499	1	-	-
o_order-priority	October 25, 2023 at 19:14:	8.45	5	15	-	-	-	-	-
o_orderdate	October 25, 2023 at 19:14:	10.00	1790	10	-	-	-	-	-
o_orderkey	October 25, 2023 at 19:14:	-	3098	-	-	12451	1	-	-
o_orderstatus	October 25, 2023 at 19:14:	1.00	3	1	-	-	-	-	-
o_ship-priority	October 25, 2023 at 19:14:	-	1	-	-	-	-	-	-
o_totalprice	October 25, 2023 at 19:14:	-	3062	-	-	422359.65	974.04	-	-

The following statistics are available:

- Column name: Column name used to generate statistics
- Last updated: Data and time when the statistics were generated
- Average length: Average length of values in the column
- Distinct values: Total number of distinct values in the column. We estimate the number of distinct values in a column with 5% relative error.
- Max value: The largest value in the column.
- Min value: The smallest value in the column.
- Max length: The length of the highest value in the column.
- Null values: The total number of null values in the column.
- True values: The total number of true values in the column.
- False values: The total number of false values in the column.

- numFiles: The total number of files in the table. This value is available under the **Advanced properties** tab.

AWS CLI

The following example shows how to retrieve column statistics using AWS CLI.

```
aws glue get-column-statistics-for-table \
    --database-name database_name \
    --table-name table_name \
    --column-names <column_name>
```

You can also view the column statistics using the GetColumnStatisticsForTable API operation.

Viewing column statistics task runs

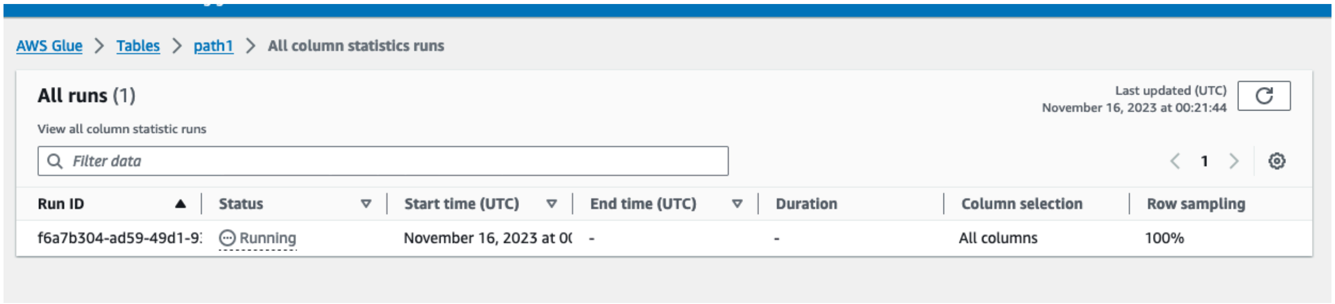
After you run a column statistics task, you can explore the task run details for a table using AWS Glue console, AWS CLI or using GetColumnStatisticsTaskRuns operation.

Console

To view column statistics task run details

1. On AWS Glue console, choose **Tables** under Data Catalog.
2. Select a table with column statistics.
3. On the **Table details** page, choose **Column statistics**.
4. Choose **View runs**.

   You can see information about all runs associated with the specified table.



All runs (1)

View all column statistic runs

Last updated (UTC)
November 16, 2023 at 00:21:44

Run ID	Status	Start time (UTC)	End time (UTC)	Duration	Column selection	Row sampling
f6a7b304-ad59-49d1-9:	Running	November 16, 2023 at 0(	-	-	All columns	100%

AWS CLI

In the following example, replace values for `DatabaseName` and `TableName` with the actual database and table name.

```
aws glue get-column-statistics-task-runs --input-cli-json file://input.json
{
    "DatabaseName": "database_name",
    "TableName": "table_name"
}
```

Stopping column statistics task run

You can stop a column statistics task run for a table using AWS Glue console, AWS CLI or using StopColumnStatisticsTaskRun operation.

Console

To stop a column statistics task run

1. On AWS Glue console, choose **Tables** under Data Catalog.
2. Select the table with the column statistics task run is in progress.
3. On the **Table details** page, choose **Column statistics**.
4. Choose **Stop**.

   If you stop the task before the run is complete, column statistics won't be generated for the table.

AWS CLI

In the following example, replace values for `DatabaseName` and `TableName` with the actual database and table name.

```
aws glue stop-column-statistics-task-run --input-cli-json file://input.json
{
    "DatabaseName": "database_name",
    "TableName": "table_name"
}
```

Deleting column statistics

You can delete column statistics using the DeleteColumnStatisticsForTable API operation or AWS CLI. The following example shows how to delete column statistics using AWS Command Line Interface (AWS CLI).

```
aws glue delete-column-statistics-for-table \
    --database-name 'database_name' \
    --table-name 'table_name' \
    --column-name 'column_name'
```

Considerations and limitations

The following considerations and limitations apply to generating column statistics.

Considerations

- Using sampling to generate statistics reduces run time, but can generate inaccurate statistics.
- Data Catalog doesn't store different versions of the statistics.
- You can only run one statistics generation task at a time per table.
- If a table is encrypted using customer AWS KMS key registered with Data Catalog, AWS Glue uses the same key to encrypt statistics.

Column statistics task supports generating statistics:

- When the IAM role has full table permissions (IAM or Lake Formation).
- When the IAM role has permissions on the table using Lake Formation hybrid access mode.

Column statistics task doesn’t support generating statistics for:

- Tables with Lake Formation cell-based access control
- Transactional data lakes - Linux foundation Delta Lake, Apache Hudi
- Tables in federated databases - Hive metastore, Amazon Redshift datashares

- Nested columns, arrays, and struct data types.
- Table that is shared with you from another account

Encrypting your Data Catalog

You can protect your metadata stored in the AWS Glue Data Catalog at rest using encryption keys managed by AWS Key Management Service (AWS KMS). You can enable Data Catalog encryption for new Data Catalog, by using the **Data Catalog settings**. You can enable or disable encryption for existing Data Catalog as needed. When enabled, AWS Glue encrypts all new metadata written to the catalog, while existing metadata remains unencrypted.

For detailed information about encrypting your Data Catalog, see Encrypting your Data Catalog.

Securing your Data Catalog using Lake Formation

AWS Lake Formation is a service that makes it easier to set up a secure data lake in AWS. It provides a central place to create and securely manage your data lakes by defining fine-granied access control permissions. Lake Formation uses the Data Catalog to store and retrieve metadata about your data lake, such as table definitions, schema information, and data access control settings.

You can register your Amazon S3 data location of the metadata table or database with Lake Formation and use it to define metadata-level permissions on the Data Catalog resources. You can also use Lake Formation to manage storage access permissions on the underlying data stored in Amazon S3 on behalf of integrated analytical engines.

For more information see What is AWS Lake Formation?.

Working with AWS Glue Data Catalog views in AWS Glue

You can create and manage views in the AWS Glue Data Catalog, commonly known as AWS Glue Data Catalog views. These views are useful because they support multiple SQL query engines, allowing you to access the same view across different AWS services, such as Amazon Athena, Amazon Redshift, and AWS Glue. You can use views based on Apache Iceberg, Apache Hudi, and Delta Lake.

By creating a view in the Data Catalog, you can use resource grants and tag-based access controls in AWS Lake Formation to grant access to it. Using this method of access control, you don't have to configure additional access to the tables referenced when creating the view. This method of granting permissions is called definer semantics, and these views are called definer views. For more information about access control in AWS Lake Formation, see [Granting and revoking permissions on Data Catalog resources] in the AWS Lake Formation Developer Guide.

Data Catalog views are useful for the following use cases:

- **Granular access control** – You can create a view that restricts data access based on the permissions the user needs. For example, you can use views in the Data Catalog to prevent employees who don't work in the HR department from seeing personally identifiable information (PII).
- **Complete view definition** – By applying filters on your view in the Data Catalog, you ensure that data records available in the view are always complete.
- **Enhanced security** – The query definition used to create the view must be complete, making Data Catalog views less susceptible to SQL commands from malicious actors.
- **Simple data sharing** – Share data with other AWS accounts without moving data, using cross-account data sharing in AWS Lake Formation.

Creating a Data Catalog view

You can create Data Catalog views using the AWS CLI and AWS Glue ETL scripts using Spark SQL. The syntax for creating a Data Catalog view includes specifying the view type as `MULTI DIALECT` and the `SECURITY` predicate as `DEFINER`, indicating a definer view.

Example SQL statement to create a Data Catalog view:

```
CREATE PROTECTED MULTI DIALECT VIEW database_name.catalog_view SECURITY DEFINER
AS SELECT order_date, sum(totalprice) AS price
FROM source_table
GROUP BY order_date;
```

After creating a Data Catalog view, you can use an IAM role with the AWS Lake Formation `SELECT` permission on the view to query it from services like Amazon Athena, Amazon Redshift, or AWS Glue ETL jobs. You don't need to grant access to the underlying tables referenced in the view.

For more information on creating and configuring Data Catalog views, see Building AWS Glue Data Catalog views in the AWS Lake Formation Developer Guide.

Supported view operations

The following command fragments show you various ways to work with Data Catalog views:

CREATE VIEW

Creates a data-catalog view. The following is a sample that shows creating a view from an existing table:

```
CREATE PROTECTED MULTI DIALECT VIEW catalog_view
SECURITY DEFINER AS SELECT * FROM my_catalog.my_database.source_table
```

ALTER VIEW

Available syntax:

```
ALTER VIEW view_name [FORCE] ADD DIALECT AS query
ALTER VIEW view_name [FORCE] UPDATE DIALECT AS query
ALTER VIEW view_name DROP DIALECT
```

You can use the `FORCE ADD DIALECT` option to force update the schema and sub objects as per the new engine dialect. Note that doing this can result in query errors if you don't also use `FORCE` to update other engine dialects. The following shows a sample:

```
ALTER VIEW catalog_view FORCE ADD DIALECTAS
SELECT order_date, sum(totalprice) AS priceFROM source_tableGROUP BY orderdate;
```

The following shows how to alter a view in order to update the dialect:

```
ALTER VIEW catalog_view UPDATE DIALECT AS
SELECT count(*) FROM my_catalog.my_database.source_table;
```

DESCRIBE VIEW

Available syntax for describing a view:

`SHOW COLUMNS {FROM|IN} view_name [{FROM|IN} database_name]` – If the user has the required AWS Glue and AWS Lake Formation permissions to describe the view, they can list the columns. The following shows a couple sample commands for showing columns:

```
SHOW COLUMNS FROM my_database.source_table;
SHOW COLUMNS IN my_database.source_table;
```

`DESCRIBE view_name` – If the user has the required AWS Glue and AWS Lake Formation permissions to describe the view, they can list the columns in the view along with its metadata.

DROP VIEW

Available syntax:

```
DROP VIEW [ IF EXISTS ] view_name
```

The following sample shows a `DROP` statement that tests if a view exists prior to dropping it:

```
DROP VIEW IF EXISTS catalog_view;
```

`SHOW CREATE VIEW view_name` – Shows the SQL statement that creates the specified view. The following is a sample that shows creating a data-catalog view:

```
SHOW CREATE TABLE my_database.catalog_view;CREATE PROTECTED MULTI DIALECT VIEW
 my_catalog.my_database.catalog_view (
  net_profit,
  customer_id,
  item_id,
  sold_date)
TBLPROPERTIES (
  'transient_lastDdlTime' = '1736267222')
SECURITY DEFINER AS SELECT * FROM
my_database.store_sales_partitioned_lf WHERE customer_id IN (SELECT customer_id from
 source_table limit 10)
```

SHOW VIEWS

List all views in the catalog, such as regular views, multi-dialect views (MDV), and MDV without Spark dialect. Available syntax is the following:

```
SHOW VIEWS [{ FROM | IN } database_name] [LIKE regex_pattern]:
```

The following shows a sample command to show views:

```
SHOW VIEWS IN marketing_analytics LIKE 'catalog_view*';
```

For more information about creating and configuring data-catalog views, see [Building AWS Glue Data Catalog views](#) in the AWS Lake Formation Developer Guide.

Querying a Data Catalog view

After creating a Data Catalog view, you can query the view. The IAM role configured in your AWS Glue jobs must have the Lake Formation **SELECT** permission on the Data Catalog view. You don't need to grant access to the underlying tables referenced in the view.

Once you have everything set up, you can query your view. For example, you can run the following query to access a view.

```
SELECT * from my_database.catalog_view LIMIT 10;
```

Limitations

Consider the following limitations when you use Data Catalog views.

- You can only create Data Catalog views with AWS Glue 5.0 and above.
- The Data Catalog view definer must have `SELECT` access to the underlying base tables accessed by the view. Creating the Data Catalog view fails if a specific base table has any Lake Formation filters imposed on the definer role.
- Base tables must not have the `IAMAllowedPrincipals` data lake permission in AWS Lake Formation. If present, the error **Multi Dialect views may only reference tables without IAMAllowedPrincipals permissions occurs**.
- The table's Amazon S3 location must be registered as a AWS Lake Formation data lake location. If the table isn't registered, the error `Multi Dialect views may only reference AWS Lake Formation managed tables` occurs. For information about how to register Amazon Amazon S3 locations in AWS Lake Formation, see [Registering an Amazon S3 location](#) in the AWS Lake Formation Developer Guide.
- You can only create `PROTECTED` Data Catalog views. `UNPROTECTED` views aren't supported.
- You can't reference tables in another AWS account in a Data Catalog view definition. You also can't reference a table in the same account that's in a separate region.
- To share data across an account or region, the entire view must be shared cross account and cross region, using AWS Lake Formation resource links.
- User-defined functions (UDFs) aren't supported.

- You can't reference other views in Data Catalog views.

Accessing the Data Catalog

You can use the AWS Glue Data Catalog (Data Catalog) to discover and understand your data. Data Catalog provides a consistent way to maintain schema definitions, data types, locations, and other metadata. You can access the Data Catalog using the following methods:

- AWS Glue console – You can access and manage the Data Catalog through the AWS Glue console, a web-based user interface. The console allows you to browse and search for databases, tables, and their associated metadata, as well as create, update, and delete metadata definitions.
- AWS Glue crawler – Crawlers are programs that automatically scan your data sources and populate the Data Catalog with metadata. You can create and run crawlers to discover and catalog data from various sources like Amazon S3, Amazon RDS, Amazon DynamoDB, Amazon CloudWatch, and JDBC-compliant relational databases such as MySQL, and PostgreSQL as well as several non-AWS sources such as Snowflake and Google BigQuery.
- AWS Glue APIs – You can access the Data Catalog programmatically using the AWS Glue APIs. These APIs allow you to interact with the Data Catalog programmatically, enabling automation and integration with other applications and services.
- AWS Command Line Interface (AWS CLI) – You can use the AWS CLI to access and manage the Data Catalog from the command line. The CLI provides commands for creating, updating, and deleting metadata definitions, as well as querying and retrieving metadata information.
- Integration with other AWS services – The Data Catalog integrates with various other AWS services, allowing you to access and utilize the metadata stored in the catalog. For example, you can use Amazon Athena to query data sources using the metadata in the Data Catalog, and use AWS Lake Formation to manage data access and governance for the Data Catalog resources.

Topics

- Connecting to the Data Catalog using AWS Glue Iceberg REST endpoint
- Connecting to the Data Catalog using AWS Glue Iceberg REST extension endpoint
- AWS Glue REST APIs for Apache Iceberg specifications
- Connecting to Data Catalog from a standalone Spark application
- Data mapping between Amazon Redshift and Apache Iceberg
- Considerations and limitations when using AWS Glue Iceberg REST Catalog APIs

Connecting to the Data Catalog using AWS Glue Iceberg REST endpoint

AWS Glue's Iceberg REST endpoint supports API operations specified in the Apache Iceberg REST specification. Using an Iceberg REST client, you can connect your application running on an analytics engine to the REST catalog hosted in the Data Catalog.

The endpoint supports both Apache Iceberg table specifications - v1 and v2, defaulting to v2. When using the Iceberg table v1 specification, you must specify v1 in the API call. Using the API operation, you can access Iceberg tables stored in both Amazon S3 object storage and Amazon S3 Table storage.

Endpoint configuration

You can access the AWS Glue Iceberg REST catalog using the service endpoint. Refer to the AWS Glue service endpoints reference guide for the region-specific endpoint. For example, when connecting to AWS Glue in the us-east-1 Region, you need to configure the endpoint URI property as follows:

```
Endpoint : https://glue.us-east-1.amazonaws.com/iceberg
```

Additional configuration properties – When using Iceberg client to connect an analytics engine like Spark to the service endpoint, you are required to specify the following application configuration properties:

```
catalog_name = "mydatacatalog"
aws_account_id = "123456789012"
aws_region = "us-east-1"
spark = SparkSession.builder \
    ... \
    .config("spark.sql.defaultCatalog", catalog_name) \
    .config(f"spark.sql.catalog.{catalog_name}",
 "org.apache.iceberg.spark.SparkCatalog") \
    .config(f"spark.sql.catalog.{catalog_name}.type", "rest") \
    .config(f"spark.sql.catalog.{catalog_name}.uri", "https://glue.
{aws_region}.amazonaws.com/iceberg") \
    .config(f"spark.sql.catalog.{catalog_name}.warehouse", "{aws_account_id}") \
    .config(f"spark.sql.catalog.{catalog_name}.rest.sigv4-enabled", "true") \
    .config(f"spark.sql.catalog.{catalog_name}.rest.signing-name", "glue") \

 .config("spark.sql.extensions","org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensi
 \
    .getOrCreate()
```

AWS Glue Iceberg endpoint `https://glue.us-east-1.amazonaws.com/iceberg` supports supports the following Iceberg REST APIs:

- GetConfig
- ListNamespaces
- CreateNamespace
- LoadNamespaceMetadata
- UpdateNamespaceProperties
- DeleteNamespace
- ListTables
- CreateTable
- LoadTable
- TableExists
- UpdateTable
- DeleteTable

Prefix and catalog path parameters

Iceberg REST catalog APIs have a free-form prefix in their request URLs. For example, the `ListNamespaces` API call uses the `GET/v1/{prefix}/namespaces` URL format. AWS Glue prefix always follows the `/catalogs/{catalog}` structure to ensure that the REST path aligns the AWS Glue multi-catalog hierarchy. The `{catalog}` path parameter can be derived based on the following rules:

Access pattern	Glue catalog ID Style	Prefix Style	Example default catalog ID	Example REST route
Access the default catalog in current account	not required	:	not applicable	GET /v1/catalogs/:/namespaces

Access pattern	Glue catalog ID Style	Prefix Style	Example default catalog ID	Example REST route
Access the default catalog in a specific account	accountID	accountID	111122223333	GET /v1/catalogs/111122223333/namespaces
Access a nested catalog in current account	catalog1/catalog2	catalog1/catalog2	rmscatalog1:db1	GET /v1/catalogs/rmscatalog1:db1/namespaces
Access a nested catalog in a specific account	accountId:catalog1/catalog2	accountId:catalog1/catalog2	123456789012/rmscatalog1:db1	GET /v1/catalogs/123456789012:rmscatalog1:db1/namespaces

This catalog ID to prefix mapping is required only when you directly call the REST APIs. When you are working with the AWS Glue Iceberg REST catalog APIs through an engine, you need to specify the AWS Glue catalog ID in the `warehouse` parameter for your Iceberg REST catalog API setting or in the `glue.id` parameter for your AWS Glue extensions API setting. For example, see how you can use it with EMR Spark in Use an Iceberg cluster with Spark.

Namespace path parameter

Namespaces in Iceberg REST catalog APIs path can have multiple levels. However, AWS Glue only supports single-level namespaces. To access a namespace in a multi-level catalog hierarchy, you can connect to a multi-level catalog above the namespace to reference the namespace. This allows any query engine that supports the 3-part notation of `catalog.namespace.table` to access objects in AWS Glue's multi-level catalog hierarchy without compatibility issues compared to using the multi-level namespace.

Connecting to the Data Catalog using AWS Glue Iceberg REST extension endpoint

AWS Glue Iceberg REST extension endpoint provides additional APIs, which are not present in the Apache Iceberg REST specification, and provides server-side scan planning capabilities. These additional APIs are used when you access tables stored in Amazon Redshift managed storage. The endpoint is accessible from an application using Apache Iceberg AWS Glue Data Catalog extensions.

Endpoint configuration – A catalog with tables in the Redshift managed storage is accessible using the service endpoint. Refer to the AWS Glue service endpoints reference guide for the region-specific endpoint. For example, when connecting to AWS Glue in the us-east-1 Region, you need to configure the endpoint URI property as follows:

```
Endpoint : https://glue.us-east-1.amazonaws.com/extensions
```

```
catalog_name = "myredshiftcatalog"
aws_account_id = "123456789012"
aws_region = "us-east-1"
spark = SparkSession.builder \
    .config("spark.sql.defaultCatalog", catalog_name) \
    .config(f"spark.sql.catalog.{catalog_name}",
 "org.apache.iceberg.spark.SparkCatalog") \
    .config(f"spark.sql.catalog.{catalog_name}.type", "glue") \
    .config(f"spark.sql.catalog.{catalog_name}.glue.id",
 "{123456789012}:redshiftnamespacecatalog/redshiftdb") \

 .config("spark.sql.extensions","org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensi
 \
    .getOrCreate()
```

AWS Glue REST APIs for Apache Iceberg specifications

This section contains specifications about the AWS Glue Iceberg REST catalog and AWS Glue extension APIs, and considerations when using these APIs.

API requests to the AWS Glue Data Catalog endpoints are authenticated using AWS Signature Version 4 (SigV4). See AWS Signature Version 4 for API requests section to learn more about AWS SigV4.

When accessing the AWS Glue service endpoint, and AWS Glue metadata, the application assumes an IAM role which requires `glue:getCatalog` IAM action.

Access to the Data Catalog, and its objects can be managed using IAM, Lake Formation, or Lake Formation hybrid mode permissions.

Federated catalogs in the Data Catalog have Lake Formation registered data locations. Lake Formation works with the Data Catalog to provide database-style permissions to manage user access to Data Catalog objects.

You can use IAM, AWS Lake Formation, or Lake Formation hybrid mode permissions to manage access to the default Data Catalog and its objects.

To create, insert, or delete data in Lake Formation managed objects, you must set up specific permissions for the IAM user or role.

- CREATE_CATALOG – Required to create catalogs
- CREATE_DATABASE – Required to create databases
- CREATE_TABLE – Required to create tables
- DELETE – Required to delete data from a table
- DESCRIBE – Required to read metadata
- DROP – Required to drop/delete a table or database
- INSERT – Needed when the principal needs to insert data into a table
- SELECT – Needed when the principal needs to select data from a table

For more information, see Lake Formation permissions reference in the AWS Lake Formation Developer Guide.

GetConfig

General information

Operation name	GetConfig
Type	Iceberg REST Catalog API
REST path	GET /iceberg/v1/config

IAM action	glue:GetCatalog
Lake Formation permissions	Not applicable
CloudTrail event	glue:GetCatalog
Open API definition	https://github.com/apache/iceberg/blob/apache-iceberg-1.6.1/open-api/rest-catalog-open-api.yaml#L67

Considerations and limitations

- The `warehouse` query parameter must be set to the AWS Glue catalog ID. If not set, the root catalog in the current account is used to return the response. For more information, see the section called "Prefix and catalog path parameters".

GetCatalog

General information

Operation name	GetCatalog
Type	AWS Glue extension API
REST path	GET/extensions/v1/catalogs/{catalog}
IAM action	glue:GetCatalog
Lake Formation permissions	DESCRIBE
CloudTrail event	glue:GetCatalog
Open API definition	https://github.com/awslabs/glue-extensions-for-iceberg/blob/main/glue-extensions-api.yaml#L40

Considerations and limitations

- The catalog path parameter must follow the style described in the the section called "Prefix and catalog path parameters" section.

ListNamespaces

General information

Operation name	ListNamespaces
Type	Iceberg REST Catalog API
REST path	GET/iceberg/v1/catalogs/{catalog}/namespaces
IAM action	glue:GetDatabase
Lake Formation permissions	ALL, DESCRIBE, SELECT
CloudTrail event	glue:GetDatabase
Open API definition	https://github.com/apache/iceberg/blob/apache-iceberg-1.6.1/open-api/rest-catalog-open-api.yaml#L205

Considerations and limitations

- The catalog path parameter must follow the style described in the the section called "Prefix and catalog path parameters" section.
- Only namespaces of the next level is displayed. To list namespaces in deeper levels, specify the nested catalog ID in the catalog path parameter.

CreateNamespace

General information

Operation name	CreateNamespace
Type	Iceberg REST Catalog API
REST path	POST/iceberg/v1/catalogs/{catalog}/namespaces
IAM action	glue:CreateDatabase
Lake Formation permissions	ALL, DESCRIBE, SELECT

CloudTrail event	glue:CreateDatabase
Open API definition	https://github.com/apache/iceberg/blob/apache-iceberg-1.6.1/open-api/rest-catalog-open-api.yaml#L256

Considerations and limitations

- The catalog path parameter must follow the style described in the the section called “Prefix and catalog path parameters” section.
- Only single level namespace can be created. To create a multi-level namespace, you must iteratively create each level, and connect to the level using the catalog path parameter.

StartCreateNamespaceTransaction

General information

Operation name	StartCreateNamespaceTransaction
Type	AWS Glue extensions API
REST path	POST/extensions/v1/catalogs/{catalog}/namespaces
IAM action	glue:CreateDatabase
Lake Formation permissions	ALL, DESCRIBE, SELECT
CloudTrail event	glue:CreateDatabase
Open API definition	https://github.com/apache/iceberg/blob/apache-iceberg-1.6.1/open-api/rest-catalog-open-api.yaml#L256

Considerations and limitations

- The catalog path parameter must follow the style described in the the section called “Prefix and catalog path parameters” section.
- You can create only single-level namespace. To create a multi-level namespaces, you must iteratively create each level, and connect to the level using the catalog path parameter.

- The API is asynchronous, and returns a transaction ID that that you can use for tracking using the `CheckTransactionStatus` API call.
- You can call this API, only if the `GetCatalog` API call contains the parameter `use-extensions=true` in the response.

LoadNamespaceMetadata

General information

Operation name	LoadNamespaceMetadata
Type	Iceberg REST Catalog API
REST path	GET/iceberg/v1/catalogs/{catalog}/namespaces/{ns}
IAM action	glue:GetDatabase
Lake Formation permissions	ALL, DESCRIBE, SELECT
CloudTrail event	glue:GetDatabase
Open API definition	https://github.com/apache/iceberg/blob/apache-iceberg-1.6.1/open-api/rest-catalog-open-api.yaml#L302

Considerations and limitations

- The catalog path parameter must follow the style described in the the section called "Prefix and catalog path parameters" section.
- You can specify only a single-level namespace in the REST path parameter. For more information, see the the section called "Namespace path parameter" section.

UpdateNamespaceProperties

General information

Operation name	UpdateNamespaceProperties
Type	Iceberg REST Catalog API

REST path	POST /iceberg/v1/catalogs/{catalog}/namespaces/{ns}/properties
IAM action	glue:UpdateDatabase
Lake Formation permissions	ALL, ALTER
CloudTrail event	glue:UpdateDatabase
Open API definition	https://github.com/apache/iceberg/blob/apache-iceberg-1.6.1/open-api/rest-catalog-open-api.yaml#L400

Considerations and limitations

- The catalog path parameter must follow the style described in the the section called "Prefix and catalog path parameters" section.
- You can specify only a single-level namespace in the REST path parameter. For more information, see the the section called "Namespace path parameter" section.

DeleteNamespace

General information

Operation name	DeleteNamespace
Type	Iceberg REST Catalog API
REST path	DELETE/iceberg/v1/catalogs/{catalog}/namespces/{ns}
IAM action	glue:DeleteDatabase
Lake Formation permissions	ALL, DROP
CloudTrail event	glue:DeleteDatabase
Open API definition	https://github.com/apache/iceberg/blob/apache-iceberg-1.6.1/open-api/rest-catalog-open-api.yaml#L365

Considerations and limitations

- The catalog path parameter must follow the style described in the the section called “Prefix and catalog path parameters” section.
- You can specify only a single-level namespace in the REST Path parameter. For more in formation, see the the section called “Namespace path parameter” section.
- If there are objects in the database, the operation will fail.
- The API is asynchronous, and returns a transaction ID that that you can use for tracking using the `CheckTransactionStatus` API call.
- The API can only be used if the `GetCatalog` API call indicates `use-extensions=true` in response.

StartDeleteNamespaceTransaction

General information

Operation name	StartDeleteNamespaceTransaction
Type	AWS Glue extensions API
REST path	DELETE /extensions/v1/catalogs/{catalog}/namespces/{ns}
IAM action	glue:DeleteDatabase
Lake Formation permissions	ALL, DROP
CloudTrail event	glue:DeleteDatabase
Open API definition	https://github.com/awslabs/glue-extensions-for-iceberg/blob/main/glue-extensions-api.yaml#L85

Considerations and limitations

- The catalog path parameter must follow the style described in the the section called “Prefix and catalog path parameters” section.
- You can specify a only single-level namespace in the REST Path parameter. For more in formation, see the the section called “Namespace path parameter” section.
- If there are objects in the database, the operation will fail.

- The API is asynchronous, and returns a transaction ID that that you can use for tracking using the `CheckTransactionStatus` API call.
- The API can only be used if the `GetCatalog` API call indicates `use-extensions=true` in response.

ListTables

General information

Operation name	ListTables
Type	Iceberg REST Catalog API
REST path	GET /iceberg/v1/catalogs/{catalog}/namespaces/{ns}/tables
IAM action	glue:GetTables
Lake Formation permissions	ALL, SELECT, DESCRIBE
CloudTrail event	glue:GetTables
Open API definition	https://github.com/apache/iceberg/blob/apache-iceberg-1.6.1/open-api/rest-catalog-open-api.yaml#L463

Considerations and limitations

- The catalog path parameter must follow the style described in the the section called “Prefix and catalog path parameters” section.
- You can specify only a single-level namespace in the REST Path parameter. For more information, see the the section called “Namespace path parameter” section.
- All tables including non-Iceberg tables will be listed. To determine if a table can be loaded as an Iceberg table or not, call `LoadTable` operation.

CreateTable

General information

Operation name	CreateTable
Type	Iceberg REST Catalog API
REST path	GET /iceberg/v1/catalogs/{catalog}/namespaces/{ns}/tables
IAM action	glue:CreateTable
Lake Formation permissions	ALL, CREATE_TABLE
CloudTrail event	glue:CreateTable
Open API definition	https://github.com/apache/iceberg/blob/apache-iceberg-1.6.1/open-api/rest-catalog-open-api.yaml#L497

Considerations and limitations

- The catalog path parameter must follow the style described in the the section called "Prefix and catalog path parameters" section.
- You can specify only a single-level namespace in the REST Path parameter. For more in formation, see the the section called "Namespace path parameter" section.
- `CreateTable` with staging is not supported. If the `stageCreate` query parameter is specified, the operation will fail.This means operation like `CREATE TABLE AS SELECT` is not supported, and you can use a combination of `CREATE TABLE` and `INSERT INTO` as a workaround.
- The `CreateTable` API operation doesn't support the option `state-create = TRUE`.

StartCreateTableTransaction

General information

Operation name	CreateTable
Type	AWS Glue extensions API

REST path	POST/extensions/v1/catalogs/{catalog}/namespaces/{ns}/tables
IAM action	glue:CreateTable
Lake Formation permissions	ALL, CREATE_TABLE
CloudTrail event	glue:CreateTable
Open API definition	https://github.com/awslabs/glue-extensions-for-iceberg/blob/main/glue-extensions-api.yaml#L107

Considerations and limitations

- The catalog path parameter must follow the style described in the the section called "Prefix and catalog path parameters" section.
- You can specify only a single-level namespace in the REST path parameter. For more in formation, see the the section called "Namespace path parameter" section.
- `CreateTable` with staging is not supported. If the `stageCreate` query parameter is specified, the operation will fail.This means operation like `CREATE TABLE AS SELECT` is not supported, and user should use a combination of `CREATE TABLE` and `INSERT INTO` to workaround.
- The API is asynchronous, and returns a transaction ID that that you can use for tracking using the `CheckTransactionStatus` API call.
- The API can only be used if the `GetCatalog` API call indicates `use-extensions=true` in response.

LoadTable

General information

Operation name	LoadTable
Type	Iceberg REST Catalog API
REST path	GET /iceberg/v1/catalogs/{catalog}/namespaces/{ns}/tables/{table}

IAM action	glue:GeTTable
Lake Formation permissions	ALL, SELECT, DESCRIBE
CloudTrail event	glue:GetTable
Open API definition	https://github.com/apache/iceberg/blob/apache-iceberg-1.6.1/open-api/rest-catalog-open-api.yaml#L616

Considerations

- The catalog path parameter must follow the style described in the the section called “Prefix and catalog path parameters” section.
- You can specify only a single level namespace in the REST Path parameter. For more in formation, see the the section called “Namespace path parameter” section.
- `CreateTable` with staging is not supported. If the `stageCreate` query parameter is specified, the operation will fail.This means operation like `CREATE TABLE AS SELECT` is not supported, and user should use a combination of `CREATE TABLE` and `INSERT INTO` to workaround.
- The API is asynchronous, and returns a transaction ID that that you can use for tracking using the `CheckTransactionStatus` API call.
- The API can only be used if the `GetCatalog` API call indicates `use-extensions=true` in response.

ExtendedLoadTable

General information

Operation name	LoadTable
Type	AWS Glue extensions API
REST path	GET /extensions/v1/catalogs/{catalog}/namespaces/{ns}/tables/{table}
IAM action	glue:GetTable
Lake Formation permissions	ALL, SELECT, DESCRIBE

CloudTrail event	glue:GetTable
Open API definition	https://github.com/awslabs/glue-extensions-for-iceberg/blob/main/glue-extensions-api.yaml#L134

Considerations

- The catalog path parameter must follow the style described in the the section called "Prefix and catalog path parameters" section.
- You can specify only a single-level namespace in the REST Path parameter. For more in formation, see the the section called "Namespace path parameter" section.
- Only `all` mode is supported for snapshots query parameter.
- Compared to `LoadTable` API, the `ExtendedLoadTable` API differs in the following ways:
  - Doesn't strictly enforce that all the fields to be available.
  - provides the following additional parameters in the config field of the response:

Additional parameters

Config key	Description
aws.server-side-capabilities.scan-planning	Indiactes if the table can be scanned using the PreparePlan table and PlanTable APIs.
aws.server-side-capabilities.data-commit	Indicates if the table can be committed using the StartUpdateTable transaction.
aws.glue.staging.location	Used for server side scan planning or data commit, a service-managed staging location that can be used for engine to write temporary data files
aws.glue.staging.access-key-id	Used for server side scan planning or data commit, a part of the temporary AWS credentials to access the service-managed staging location
aws.glue.staging.secret-access-key	Used for server side scan planning or data commit, a part of the temporary AWS credentials to access the service-managed staging location.

Config key	Description
aws.glue.staging.session-token	Used for server side scan planning or data commit, a part of the temporary AWS credentials to access the service-managed staging location.
aws.glue.staging.expiration-ms	Used for server side scan planning or data commit, expiration time of the credentials to access the service-managed staging. location.
aws.glue.staging.data-transfer-role-arn	Used for server side scan planning or data commit, an IAM role that can be assumed to taccess the service-managed staging location.

PreplanTable

General information

Operation name	PreplanTable
Type	AWS Glue extensions API
REST path	POST /extensions/v1/catalogs/{catalog}/namespaces/{ns}/tables/{table}/preplan
IAM action	glue:GetTable
Lake Formation permissions	ALL, SELECT, DESCRIBE
CloudTrail event	glue:GetTable
Open API definition	https://github.com/awslabs/glue-extensions-for-iceberg/blob/main/glue-extensions-api.yaml#L211

Considerations

- The catalog path parameter should follow the style described in the the section called “Prefix and catalog path parameters” section.

- You can specify only a single-level namespace in the REST Path parameter. For more in formation, see the the section called "Namespace path parameter" section.
- Caller of this API should always determine if there are remaining results to fetch based on the page token. A response with empty page item but a pagination token is possible if the server side is still processing but is not able to produce any result in the given response time.
- You can use this API only if the `ExtendedLoadTable` API response contains `aws.server-side-capabilities.scan-planning=true`.

PlanTable

General information

Operation name	PlanTable
Type	AWS Glue extensions API
REST path	POST /extensions/v1/catalogs/{catalog}/namespaces/{ns}/tables/{table}/plan
IAM action	glue:GetTable
Lake Formation permissions	ALL, SELECT, DESCRIBE
CloudTrail event	glue:GetTable
Open API definition	https://github.com/awslabs/glue-extensions-for-iceberg/blob/main/glue-extensions-api.yaml#L243

Considerations

- The catalog path parameter must follow the style described in the the section called "Prefix and catalog path parameters" section.
- You can specify only a single-level namespace in the REST Path parameter. For more in formation, see the the section called "Namespace path parameter" section.
- Caller of this API should always determine if there are remaining results to fetch based on the page token. A response with empty page item but a pagination token is possible if the server side is still processing but is not able to produce any result in the given response time.

- You can use this API only if the `ExtendedLoadTable` API response contains `aws.server-side-capabilities.scan-planning=true`.

TableExists

General information

Operation name	TableExists
Type	Iceberg REST Catalog API
REST path	HEAD/iceberg/v1/catalogs/{catalog}/namespaces/{ns}/tables/{table}
IAM action	glue:GetTable
Lake Formation permissions	ALL, SELECT, DESCRIBE
CloudTrail event	glue:GetTable
Open API definition	https://github.com/apache/iceberg/blob/apache-iceberg-1.6.1/open-api/rest-catalog-open-api.yaml#L833

Considerations

- The catalog path parameter must follow the style described in the the section called "Prefix and catalog path parameters" section.
- You can specify only a single-level namespace in the REST Path parameter. For more in formation, see the the section called "Namespace path parameter" section.

UpdateTable

General information

Operation name	UpdateTable
Type	Iceberg REST Catalog API

REST path	POST /iceberg/v1/catalogs/{catalog}/namespaces/{ns}/tables/{table}
IAM action	glue:UpdateTable
Lake Formation permissions	ALL, ALTER
CloudTrail event	glue:UpdateTable
Open API definition	https://github.com/apache/iceberg/blob/apache-iceberg-1.6.1/open-api/rest-catalog-open-api.yaml#L677

Considerations

- The catalog path parameter must follow the style described in the the section called “Prefix and catalog path parameters” section.
- You can specify only a single-level namespace in the REST Path parameter. For more in formation, see the the section called “Namespace path parameter” section.

StartUpdateTableTransaction

General information

Operation name	StartUpdateTableTransaction
Type	AWS Glue extension API
REST path	POST/extensions/v1/catalogs/{catalog}/namespaces/{ns}/tables/{table}
IAM action	glue:UpdateTable
Lake Formation permissions	ALL, ALTER
CloudTrail event	glue:UpdateTable
Open API definition	https://github.com/awslabs/glue-extensions-for-iceberg/blob/main/glue-extensions-api.yaml#L154

Considerations

- The catalog path parameter must follow the style described in the the section called “Prefix and catalog path parameters” section.
- You can specify only a single-level namespace in the REST Path parameter. For more in formation, see the the section called “Namespace path parameter” section.
- The API is asynchronous, and returns a transaction ID that that you can use for tracking using the `CheckTransactionStatus` API call.
- A `RenamTable` operation can also be performed through this API. When that happens, the caller must also ahve glue:CreateTable or LakeFormation CREATE_TABLE permission for the table to be renamed to.
- You can use this API only if the `ExtendedLoadTable` API response contains `aws.server-side-capabilities.scan-planning=true`.

DeleteTable

General information

Operation name	DeleteTable
Type	Iceberg REST Catalog API
REST path	DELETE/iceberg/v1/catalogs/{catalog}/namespaces/{ns}/tables/{table}
IAM action	glue:DeleteTable
Lake Formation permissions	ALL, DROP
CloudTrail event	glue:DeleteTable
Open API definition	https://github.com/apache/iceberg/blob/apache-iceberg-1.6.1/open-api/rest-catalog-open-api.yaml#L793

Considerations

- The catalog path parameter should follow the style described in the the section called “Prefix and catalog path parameters” section.

- You can specify only a single-level namespace in the REST Path parameter. For more in formation, see the the section called “Namespace path parameter” section.
- `DeleteTable` API operation supports a purge option. When purge is set to `true`, the table data is deleted, otherwise data is not deleted. For tables in Amazon S3, the operation does not delete table data. The operation fails when table is stored in Amazon S3, and `purge = TRUE, .`

  For tables stored in Amazon Redshift managed storage, the operation will delete table data, similar to `DROP TABLE`behavior in Amazon Redshift. The operation fails when table is stored in Amazon Redshift and `purge = FALSE`.
- `purgeRequest=true` is not supported.

StartDeleteTableTransaction

General information

Operation name	StartDeleteTableTransaction
Type	AWS Glue extensions API
REST path	DELETE /extensions/v1/catalogs/{catalog}/namespaces/{ns}/tables/{table}
IAM action	glue:DeleteTable
Lake Formation permissions	ALL, DROP
CloudTrail event	glue:DeleteTable
Open API definition	https://github.com/apache/iceberg/blob/apache-iceberg-1.6.1/open-api/rest-catalog-open-api.yaml#L793

Considerations

- The catalog path parameter must follow the style described in the the section called “Prefix and catalog path parameters” section.
- You can specify only a single-level namespace in the REST Path parameter. For more in formation, see the the section called “Namespace path parameter” section.
- `purgeRequest=false` is not supported.

- The API is asynchronous, and returns a transaction ID that can be tracked through `CheckTransactionStatus`.

CheckTransactionStatus

General information

Operation name	CheckTransactionStatus
Type	AWS Glue extensions API
REST path	POST/extensions/v1/transactions/status
IAM action	The same permission as the action that initiates the transaction
Lake Formation permissions	The same permission as the action that initiates the transaction
Open API definition	https://github.com/awslabs/glue-extensions-for-iceberg/blob/main/glue-extensions-api.yaml#L273

Considerations

- The catalog path parameter must follow the style described in the the section called "Prefix and catalog path parameters" section.

Connecting to Data Catalog from a standalone Spark application

You can connect to the Data Catalog from a stand application using an Apache Iceberg connector.

1. Create an IAM role for Spark application.
2. Connect to AWS Glue Iceberg Rest endpoint using Iceberg connector.

```
# configure your application. Refer to https://docs.aws.amazon.com/cli/latest/
userguide/cli-configure-envvars.html for best practices on configuring environment
 variables.
export AWS_ACCESS_KEY_ID=$(aws configure get appUser.aws_access_key_id)
export AWS_SECRET_ACCESS_KEY=$(aws configure get appUser.aws_secret_access_key)
export AWS_SESSION_TOKEN=$(aws configure get appUser.aws_secret_token)

export AWS_REGION=us-east-1
export REGION=us-east-1
export AWS_ACCOUNT_ID = {specify your aws account id here}

~/spark-3.5.3-bin-hadoop3/bin/spark-shell \
    --packages org.apache.iceberg:iceberg-spark-runtime-3.4_2.12:1.6.0 \
    --conf
 "spark.sql.extensions=org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions"
 \
    --conf "spark.sql.defaultCatalog=spark_catalog" \
    --conf "spark.sql.catalog.spark_catalog=org.apache.iceberg.spark.SparkCatalog" \
    --conf "spark.sql.catalog.spark_catalog.type=rest" \
    --conf "spark.sql.catalog.spark_catalog.uri=https://glue.us-east-1.amazonaws.com/
iceberg" \
    --conf "spark.sql.catalog.spark_catalog.warehouse = {AWS_ACCOUNT_ID}" \
    --conf "spark.sql.catalog.spark_catalog.rest.sigv4-enabled=true" \
    --conf "spark.sql.catalog.spark_catalog.rest.signing-name=glue" \
    --conf "spark.sql.catalog.spark_catalog.rest.signing-region=us-east-1" \
    --conf "spark.sql.catalog.spark_catalog.io-
impl=org.apache.iceberg.aws.s3.S3FileIO" \
    --conf
 "spark.hadoop.fs.s3a.aws.credentials.provider=org.apache.hadoop.fs.s3a.SimpleAWSCredentialP
```

3. Query data in the Data Catalog.

```
spark.sql("create database myicebergdb").show()
spark.sql("""CREATE TABLE myicebergdb.mytbl (name string) USING iceberg location
 's3://bucket_name/mytbl'""")
spark.sql("insert into myicebergdb.mytbl values('demo') ").show()
```

Data mapping between Amazon Redshift and Apache Iceberg

Redshift and Iceberg support various data types. The following compatibility matrix outlines the support and limitations when mapping data between these two data systems. Please refer to Amazon Redshift Data Types and Apache Iceberg Table Specifications for more details on supported data types in respective data systems.

Redshift data type	Aliases	Iceberg data type
SMALLINT	INT2	int
INTEGER	INT, INT4	int
BIGINT	INT8	long
DECIMAL	NUMERIC	decimal
REAL	FLOAT4	float
REAL	FLOAT4	float
DOUBLE PRECISION	FLOAT8, FLOAT	double
CHAR	CHARACTER, NCHAR	string
VARCHAR	CHARACTER VARYING, NVARCHAR	string
BPCHAR		string
TEXT		string
DATE		date
TIME	TIME WITHOUT TIMEZONE	time
TIME	TIME WITH TIMEZONE	not supported
TIMESTAMP	TIMESTAMP WITHOUT TIMEZONE	TIMESTAMP
TIMESTAMPZ	TIMESTAMP WITH TIMEZONE	TIMESTAMPZ
INTERVAL YEAR TO MONTH		Not supported
INTERVAL DAY TO SECOND		Not supported
BOOLEAN	BOOL	bool

Redshift data type	Aliases	Iceberg data type
HLLSKETCH		Not supported
SUPER		Not supported
VARBYTE	VARBINARY, BINARY VARYING	binary
GEOMETRY		Not supported
GEOGRAPHY		Not supported

Considerations and limitations when using AWS Glue Iceberg REST Catalog APIs

Following are the considerations and limitations when using the Apache Iceberg REST Catalog Data Definition Language (DDL) operation behavior.

Considerations

- **`RenameTable` API behavior –** The `RenameTable` operation is supported in tables in Amazon Redshift but not in Amazon S3.
- **DDL operations for namespaces and tables in Amazon Redshift –** Create, Update, Delete operations for namespaces and tables in Amazon Redshift are asynchronous operations because they are dependent on when Amazon Redshift managed workgroup is available and whether a conflicting DDL and DML transaction is in progress and operation has to wait for lock and then attempt to commit changes.

Limitations

- View APIs in the Apache Iceberg REST specification are not supported in AWS Glue Iceberg REST Catalog.

AWS Glue Data Catalog best practices

This section covers best practices for effectively managing and utilizing the AWS Glue Data Catalog. It emphasizes practices such as efficient crawler usage, metadata organization, security, performance optimization, automation, data governance, and integration with other AWS services.

- **Use crawlers effectively –** Run crawlers regularly to keep the Data Catalog up-to-date with changes in your data sources. Use incremental crawls for frequently changing data sources to improve performance. Configure crawlers to automatically add new partitions or update schemas when changes are detected.
- **Organize and name metadata tables –** Establish a consistent naming convention for databases and tables in the Data Catalog. Group related data sources into logical databases or folders for better organization. Use descriptive names that convey the purpose and content of each table.
- **Manage schemas effectively** – Take advantage of the schema inference capabilities of AWS Glue crawlers. Review and update schema changes before applying them to avoid breaking downstream applications. Use schema evolution features to handle schema changes gracefully.
- **Secure the Data Catalog –** Enable data encryption at rest and in transit for the Data Catalog. Implement fine-grained access control policies to restrict access to sensitive data. Regularly audit and review Data Catalog permissions and activity logs.
- **Integrate with other AWS services** Data Catalog Use the Data Catalog as a centralized metadata layer for services like Amazon Athena, Redshift Spectrum, and AWS Lake Formation. Leverage AWS Glue ETL jobs to transform and load data into various data stores while maintaining metadata in the Data Catalog.
- **Monitor and optimize performance** Data Catalog Monitor the performance of crawlers and ETL jobs using Amazon CloudWatch metrics. Partition large datasets in the Data Catalog to improve query performance. Implement performance optimizations for frequently accessed metadata.
- **Stay updated with AWS Glue documentation and best practices** Data Catalog Regularly check the AWS Glue documentation and AWS Glue resources for the latest updates, best practices, and recommendations. Attend AWS Glue webinars, workshops, and other events to learn from experts and stay informed about new features and capabilities.

Monitoring Data Catalog usage metrics in Amazon CloudWatch

AWS Glue Data Catalog usage metrics are now available with Amazon CloudWatch, simplifying the monitoring and understanding of resource utilization in your Data Catalog. You now have immediate visibility into your Glue Catalog API usage of catalogs, databases, tables, partitions, and connections, making it easier to maintain oversight of your Data Catalog.

Overview of Data Catalog metrics

AWS Glue Data Catalog automatically publishes usage metrics to Amazon CloudWatch. With CloudWatch metrics integration, you can track critical performance indicators every minute, including:

- Table requests
- Partition indexes created
- Connections updated
- Statistics updated

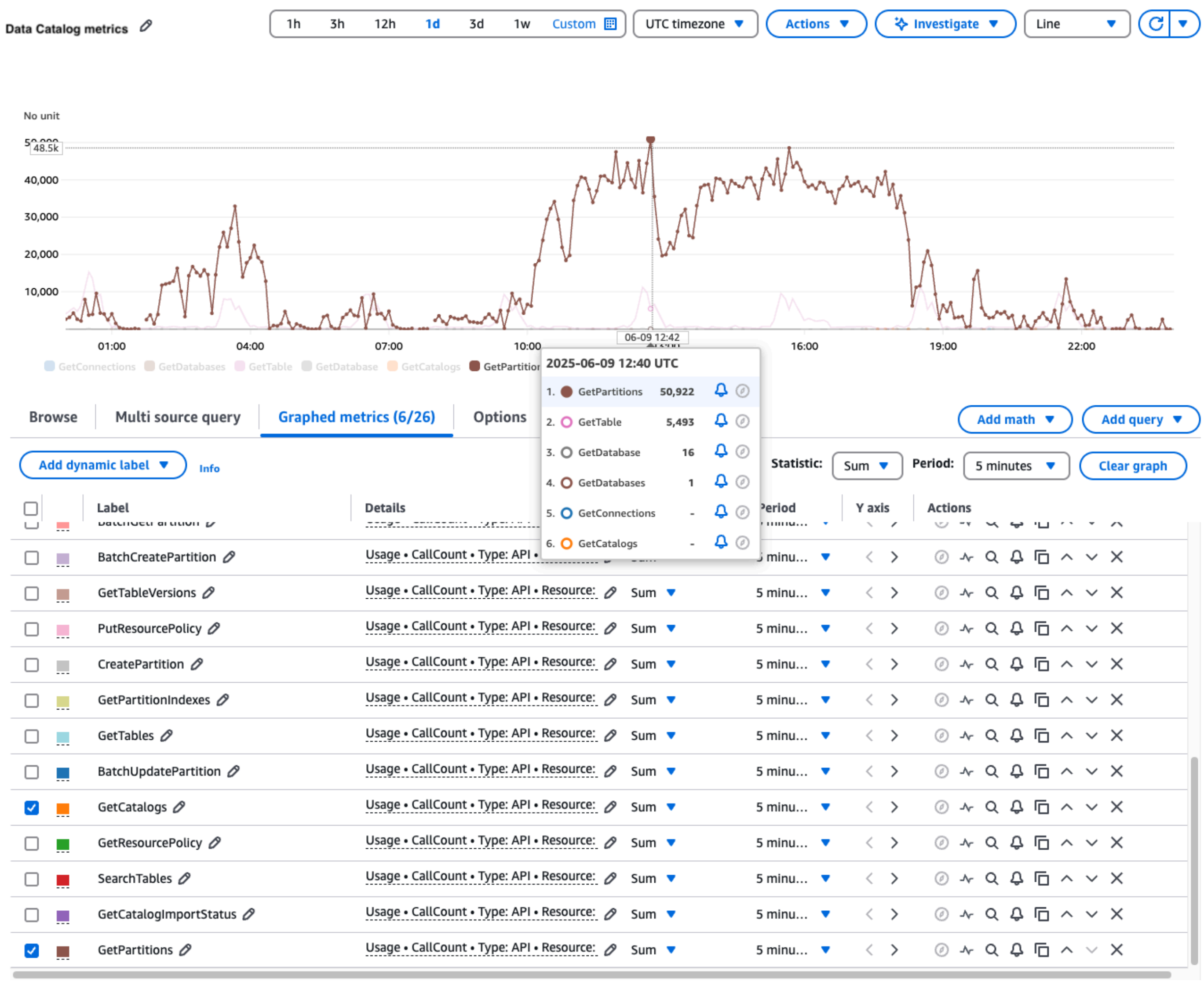
These metrics help you identify bottlenecks, detect anomalies, and make data-driven decisions to improve overall data catalog reliability. You can also set up CloudWatch alarms to receive notifications when metrics exceed specified thresholds, allowing for proactive management of your deployment.

Adding metrics to your CloudWatch dashboard

You can create custom dashboards to monitor your AWS Glue Data Catalog resources and set up alarms to be notified of any unusual activity.

You can add Data Catalog metrics to your CloudWatch dashboard by following these steps:

1. Open the CloudWatch console at https://console.aws.amazon.com/cloudwatch/.
2. In the navigation pane, choose **Metrics**.
3. Choose **All metrics**.
4. Choose **Usage>By AWS resource**.
5. Filter by **Glue** to see available metrics.
6. Select the metrics you want to add to your dashboard.
7. Add metrics for catalogs, databases, tables, partitions, and connections to your CloudWatch graph.


Data Catalog metrics
1h 3h 12h 1d 3d 1w Custom
UTC timezone
Actions
Investigate
Line
No unit
48.5k
40,000
30,000
20,000
10,000
01:00
04:00
07:00
10:00
06-09 12:42
16:00
19:00
22:00
GetConnections GetDatabases GetTable GetDatabase GetCatalogs GetPartitions
Browse
Multi source query
Graphed metrics (6/26)
Options
Add math
Add query
Add dynamic label
Info
Clear graph

2025-06-09 12:40 UTC

1.	GetPartitions	50,922
2.	GetTable	5,493
3.	GetDatabase	16
4.	GetDatabases	1
5.	GetConnections	-
6.	GetCatalogs	-

Statistic: Sum Period: 5 minutes

☐	Label	Details	Statistic	Period	Y axis	Actions
☐	BatchCreatePartition	Usage • CallCount • Type: API •		5 minu...		
☐	GetTableVersions	Usage • CallCount • Type: API • Resource:	Sum	5 minu...		
☐	PutResourcePolicy	Usage • CallCount • Type: API • Resource:	Sum	5 minu...		
☐	CreatePartition	Usage • CallCount • Type: API • Resource:	Sum	5 minu...		
☐	GetPartitionIndexes	Usage • CallCount • Type: API • Resource:	Sum	5 minu...		
☐	GetTables	Usage • CallCount • Type: API • Resource:	Sum	5 minu...		
☐	BatchUpdatePartition	Usage • CallCount • Type: API • Resource:	Sum	5 minu...		
☑	GetCatalogs	Usage • CallCount • Type: API • Resource:	Sum	5 minu...		
☐	GetResourcePolicy	Usage • CallCount • Type: API • Resource:	Sum	5 minu...		
☐	SearchTables	Usage • CallCount • Type: API • Resource:	Sum	5 minu...		
☐	GetCatalogImportStatus	Usage • CallCount • Type: API • Resource:	Sum	5 minu...		
☑	GetPartitions	Usage • CallCount • Type: API • Resource:	Sum	5 minu...		

You can configure custom alarms that trigger automatically when API usage exceeds your defined thresholds to identify abnormalities in your data catalog usage.

For detailed instructions on setting up alarms, see Creating a Metrics Insights CloudWatch alarm.

AWS Glue Schema registry

> **Note**
>
> AWS Glue Schema Registry is not supported in the following Regions in the AWS Glue console: Asia Pacific (Jakarta),, Europe (Zurich) and Middle East (UAE).

The AWS Glue Schema registry allows you to centrally discover, control, and evolve data stream schemas. A *schema* defines the structure and format of a data record. With AWS Glue Schema registry, you can manage and enforce schemas on your data streaming applications using convenient integrations with Apache Kafka, Amazon Managed Streaming for Apache Kafka, Amazon Kinesis Data Streams, Amazon Managed Service for Apache Flink, and AWS Lambda.

The Schema registry supports AVRO (v1.11.4) data format, JSON Data format with JSON Schema format for the schema (specifications Draft-04, Draft-06, and Draft-07) with JSON schema validation using the Everit library, Protocol Buffers (Protobuf) versions proto2 and proto3 without support for `extensions` or `groups`, and Java language support, with other data formats and languages to come. Supported features include compatibility, schema sourcing via metadata, auto-registration of schemas, IAM compatibility, and optional ZLIB compression to reduce storage and data transfer. The Schema registry is serverless and free to use.

Using a schema as a data format contract between producers and consumers leads to improved data governance, higher quality data, and enables data consumers to be resilient to compatible upstream changes.

The Schema registry allows disparate systems to share a schema for serialization and de-serialization. For example, assume you have a producer and consumer of data. The producer knows the schema when it publishes the data. The Schema Registry supplies a serializer and deserializer for certain systems such as Amazon MSK or Apache Kafka.

For more information, see How the schema registry works.

Topics

- Schemas
- Registries
- Schema versioning and compatibility
- Open source Serde libraries
- Quotas of the Schema Registry
- How the schema registry works
- Getting started with schema registry
- Integrating with AWS Glue Schema Registry
- Migration from a third-party schema registry to AWS Glue Schema Registry

Schemas

A *schema* defines the structure and format of a data record. A schema is a versioned specification for reliable data publication, consumption, or storage.

In this example schema for Avro, the format and structure are defined by the layout and field names, and the format of the field names is defined by the data types (e.g., `string`, `int`).

```
{
    "type": "record",
    "namespace": "ABC_Organization",
    "name": "Employee",
    "fields": [
        {
            "name": "Name",
            "type": "string"
        },
        {
            "name": "Age",
            "type": "int"
        },
        {
            "name": "address",
            "type": {
                "type": "record",
                "name": "addressRecord",
                "fields": [
                    {
                        "name": "street",
                        "type": "string"
                    },
                    {
                        "name": "zipcode",
                        "type": "int"
                    }
                ]
            }
        }
    ]
}
```

In this example JSON schema draft-07 for JSON, the format is defined by the JSON Schema organization.

```
{
 "$id": "https://example.com/person.schema.json",
 "$schema": "http://json-schema.org/draft-07/schema#",
 "title": "Person",
 "type": "object",
 "properties": {
  "firstName": {
   "type": "string",
   "description": "The person's first name."
  },
  "lastName": {
   "type": "string",
   "description": "The person's last name."
  },
  "age": {
   "description": "Age in years which must be equal to or greater than zero.",
   "type": "integer",
   "minimum": 0
  }
 }
}
```

In this example for Protobuf, the format is defined by the version 2 of the Protocol Buffers language (proto2).

```
syntax = "proto2";

package tutorial;

option java_multiple_files = true;
option java_package = "com.example.tutorial.protos";
option java_outer_classname = "AddressBookProtos";

message Person {
  optional string name = 1;
  optional int32 id = 2;
  optional string email = 3;

  enum PhoneType {
    MOBILE = 0;
    HOME = 1;
    WORK = 2;
  }

  message PhoneNumber {
    optional string number = 1;
    optional PhoneType type = 2 [default = HOME];
  }

  repeated PhoneNumber phones = 4;
}

message AddressBook {
  repeated Person people = 1;
}
```

Registries

A *registry* is a logical container of schemas. Registries allow you to organize your schemas, as well as manage access control for your applications. A registry has an Amazon Resource Name (ARN) to allow you to organize and set different access permissions to schema operations within the registry.

You may use the default registry or create as many new registries as necessary.

AWS Glue Schema Registry Hierarchy

- RegistryName: [string]
  - RegistryArn: [AWS ARN]
  - CreatedTime: [timestamp]
  - UpdatedTime: [timestamp]
- SchemaName: [string]
  - SchemaArn: [AWS ARN]
  - DataFormat: [Avro, Json, or Protobuf]
  - Compatibility: [eg. BACKWARD, BACKWARD_ALL, FORWARD, FORWARD_ALL, FULL, FULL_ALL, NONE, DISABLED]
  - Status: [eg. PENDING, AVAILABLE, DELETING]
  - SchemaCheckpoint: [integer]

  - CreatedTime: [timestamp]
  - UpdatedTime: [timestamp]

- SchemaVersion: [string]
  - SchemaVersionNumber: [integer]
  - Status: [eg. PENDING, AVAILABLE, DELETING, FAILURE]
  - SchemaDefinition: [string, Value: JSON]
  - CreatedTime: [timestamp]

- SchemaVersionMetadata: [list]
  - MetadataKey: [string]
  - MetadataInfo
  - MetadataValue: [string]
  - CreatedTime: [timestamp]

Schema versioning and compatibility

Each schema can have multiple versions. Versioning is governed by a compatibility rule that is applied on a schema. Requests to register new schema versions are checked against this rule by the Schema Registry before they can succeed.

A schema version that is marked as a checkpoint is used to determine the compatibility of registering new versions of a schema. When a schema first gets created the default checkpoint will be the first version. As the schema evolves with more versions, you can use the CLI/SDK to change the checkpoint to a version of a schema using the `UpdateSchema` API that adheres to a set of constraints. In the console, editing the schema definition or compatibility mode will change the checkpoint to the latest version by default.

Compatibility modes allow you to control how schemas can or cannot evolve over time. These modes form the contract between applications producing and consuming data. When a new version of a schema is submitted to the registry, the compatibility rule applied to the schema name is used to determine if the new version can be accepted. There are 8 compatibility modes: NONE, DISABLED, BACKWARD, BACKWARD_ALL, FORWARD, FORWARD_ALL, FULL, FULL_ALL.

In the Avro data format, fields may be optional or required. An optional field is one in which the `Type` includes null. Required fields do not have null as the `Type`.

In the Protobuf data format, fields can be optional (including repeated) or required in proto2 syntax, while all fields are optional (including repeated) in proto3 syntax. All compatibility rules are determined based on the understanding of the Protocol Buffers specifications as well as the guidance from the Google Protocol Buffers documentation.

- *NONE*: No compatibility mode applies. You can use this choice in development scenarios or if you do not know the compatibility modes that you want to apply to schemas. Any new version added will be accepted without undergoing a compatibility check.
- *DISABLED*: This compatibility choice prevents versioning for a particular schema. No new versions can be added.
- *BACKWARD*: This compatibility choice is recommended because it allows consumers to read both the current and the previous schema version. You can use this choice to check compatibility against the previous schema version when you delete fields or add optional fields. A typical use case for BACKWARD is when your application has been created for the most recent schema.

  AVRO

  For example, assume you have a schema defined by first name (required), last name (required), email (required), and phone number (optional).

  If your next schema version removes the required email field, this would successfully register. BACKWARD compatibility requires consumers to be able to read the current and previous schema version. Your consumers will be able to read the new schema as the extra email field from old messages is ignored.

  If you have a proposed new schema version that adds a required field, for example, zip code, this would not successfully register with BACKWARD compatibility. Your consumers on the new version would not be able to read old messages before the schema change, as they are missing the required zip code field. However, if the zip code field was set as optional in the new schema, then the proposed version would successfully register as consumers can read the old schema without the optional zip code field.

  JSON

  For example, assume you have a schema version defined by first name (optional), last name (optional), email (optional) and phone number (optional).

  If your next schema version adds the optional phone number property, this would successfully register as long as the original schema version does not allow any additional properties by setting the `additionalProperties` field to false. BACKWARD compatibility requires consumers to be able to read the current and previous schema version. Your consumers will be able to read data produced with the original schema where phone number property does not exist.

If you have a proposed new schema version that adds the optional phone number property, this would not successfully register with BACKWARD compatibility when the original schema version sets the `additionalProperties` field to true, namely allowing any additional property. Your consumers on the new version would not be able to read old messages before the schema change, as they cannot read data with phone number property in a different type, for example string instead of number.

PROTOBUF

For example, assume you have a schema version defined by a Message `Person` with `first name` (required), `last name` (required), `email` (required), and `phone number` (optional) fields under proto2 syntax.

Similar to AVRO scenarios, if your next schema version removes the required `email` field, this would successfully register. BACKWARD compatibility requires consumers to be able to read the current and previous schema version. Your consumers will be able to read the new schema as the extra `email` field from old messages is ignored.

If you have a proposed new schema version that adds a required field, for example, `zip code`, this would not successfully register with BACKWARD compatibility. Your consumers on the new version would not be able to read old messages before the schema change, as they are missing the required `zip code` field. However, if the `zip code` field was set as optional in the new schema, then the proposed version would successfully register as consumers can read the old schema without the optional `zip code` field.

In case of a gRPC use case, adding new RPC service or RPC method is a backward compatible change. For example, assume you have a schema version defined by an RPC service `MyService` with two RPC methods `Foo` and `Bar`.

If your next schema version adds a new RPC method called `Baz`, this would successfully register. Your consumers will be able to read data produced with the original schema according to BACKWARD compatibility since the newly added RPC method `Baz` is optional.

If you have a proposed new schema version that removes the existing RPC method `Foo`, this would not successfully register with BACKWARD compatibility. Your consumers on the new version would not be able to read old messages before the schema change, as they cannot understand and read data with the non-existent RPC method `Foo` in a gRPC application.

- *BACKWARD_ALL*: This compatibility choice allows consumers to read both the current and all previous schema versions. You can use this choice to check compatibility against all previous schema versions when you delete fields or add optional fields.
- *FORWARD*: This compatibility choice allows consumers to read both the current and the subsequent schema versions, but not necessarily later versions. You can use this choice to check compatibility against the last schema version when you add fields or delete optional fields. A typical use case for FORWARD is when your application has been created for a previous schema and should be able to process a more recent schema.

AVRO

For example, assume you have a schema version defined by first name (required), last name (required), email (optional).

If you have a new schema version that adds a required field, e.g. phone number, this would successfully register. FORWARD compatibility requires consumers to be able to read data produced with the new schema by using the previous version.

If you have a proposed schema version that deletes the required first name field, this would not successfully register with FORWARD compatibility. Your consumers on the prior version would not be able to read the proposed schemas as they are missing the required first name field. However, if the first name field was originally optional, then the proposed new schema would successfully register as the consumers can read data based on the new schema that doesn't have the optional first name field.

JSON

For example, assume you have a schema version defined by first name (optional), last name (optional), email (optional) and phone number (optional).

If you have a new schema version that removes the optional phone number property, this would successfully register as long as the new schema version does not allow any additional properties by setting the `additionalProperties` field to false. FORWARD compatibility requires consumers to be able to read data produced with the new schema by using the previous version.

If you have a proposed schema version that deletes the optional phone number property, this would not successfully register with FORWARD compatibility when the new schema version sets the `additionalProperties` field to true, namely allowing any additional property. Your consumers on the prior version would not be able to read the proposed schemas as they could have phone number property in a different type, for example string instead of number.

PROTOBUF

For example, assume you have a schema version defined by a Message `Person` with `first name` (required), `last name` (required), `email` (optional) fields under proto2 syntax.

Similar to AVRO scenarios, if you have a new schema version that adds a required field, e.g. `phone number`, this would successfully register. FORWARD compatibility requires consumers to be able to read data produced with the new schema by using the previous version.

If you have a proposed schema version that deletes the required `first name` field, this would not successfully register with FORWARD compatibility. Your consumers on the prior version would not be able to read the proposed schemas as they are missing the required `first name` field. However, if the `first name` field was originally optional, then the proposed new schema would successfully register as the consumers can read data based on the new schema that doesn’t have the optional `first name` field.

In case of a gRPC use case, removing an RPC service or RPC method is a forward-compatible change. For example, assume you have a schema version defined by an RPC service `MyService` with two RPC methods `Foo` and `Bar`.

If your next schema version deletes the existing RPC method named `Foo`, this would successfully register according to FORWARD compatibility as the consumers can read data produced with the new schema by using the previous version. If you have a proposed new schema version that adds an RPC method `Baz`, this would not successfully register with FORWARD compatibility. Your consumers on the prior version would not be able to read the proposed schemas as they are missing the RPC method `Baz`.

- *FORWARD_ALL*: This compatibility choice allows consumers to read data written by producers of any new registered schema. You can use this choice when you need to add fields or delete optional fields, and check compatibility against all previous schema versions.

- *FULL*: This compatibility choice allows consumers to read data written by producers using the previous or next version of the schema, but not earlier or later versions. You can use this choice to check compatibility against the last schema version when you add or remove optional fields.
- *FULL_ALL*: This compatibility choice allows consumers to read data written by producers using all previous schema versions. You can use this choice to check compatibility against all previous schema versions when you add or remove optional fields.

Open source Serde libraries

AWS provides open-source Serde libraries as a framework for serializing and deserializing data. The open source design of these libraries allows common open-source applications and frameworks to support these libraries in their projects.

For more details on how the Serde libraries work, see How the schema registry works.

Quotas of the Schema Registry

Quotas, also referred to as limits in AWS, are the maximum values for the resources, actions, and items in your AWS account. The following are soft limits for the Schema Registry in AWS Glue.

Schema version metadata key-value pairs

You can have up to 10 key-value pairs per SchemaVersion per AWS Region.

You can view or set the key-value metadata pairs using the QuerySchemaVersionMetadata action (Python: query_schema_version_metadata) or PutSchemaVersionMetadata action (Python: put_schema_version_metadata) APIs.

The following are hard limits for the Schema Registry in AWS Glue.

Registries

You can have up to 100 registries per AWS Region for this account.

SchemaVersion

You can have up to 10000 schema versions per AWS Region for this account.

Each new schema creates a new schema version, so you can theoretically have up to 10000 schemas per account per region, if each schema has only one version.

Schema payloads

There is a size limit of 170KB for schema payloads.

How the schema registry works

This section describes how the serialization and deserialization processes in schema registry work.

1. Register a schema: If the schema doesn't already exist in the registry, the schema can be registered with a schema name equal to the name of the destination (e.g., test_topic, test_stream, prod_firehose) or the producer can provide a custom name for the schema. Producers can also add key-value pairs to the schema as metadata, such as source: msk_kafka_topic_A, or apply AWS tags to schemas on schema creation. Once a schema is registered the Schema Registry returns the schema version ID to the serializer. If the schema exists but the serializer is using a new version that doesn't exist, the Schema Registry will check the schema reference a compatibility rule to ensure the new version is compatible before registering it as a new version.

   There are two methods of registering a schema: manual registration and auto-registration. You can register a schema manually via the AWS Glue console or CLI/SDK.

   When auto-registration is turned on in the serializer settings, automatic registration of the schema will be performed. If `REGISTRY_NAME` is not provided in the producer configurations, then auto-registration will register the new schema version under the default registry (default-registry). See Installing SerDe Libraries for information on specifying the auto-registration property.
2. Serializer validates data records against the schema: When the application producing data has registered its schema, the Schema Registry serializer validates the record being produced by the application is structured with the fields and data types matching a registered schema. If the schema of the record does not match a registered schema, the serializer will return an exception and the application will fail to deliver the record to the destination.

   If no schema exists and if the schema name is not provided via the producer configurations, then the schema is created with the same name as the topic name (if Apache Kafka or Amazon MSK) or stream name (if Kinesis Data Streams).

   Every record has a schema definition and data. The schema definition is queried against the existing schemas and versions in the Schema Registry.

   By default, producers cache schema definitions and schema version IDs of registered schemas. If a record's schema version definition does not match what's available in cache, the producer will attempt to validate the schema with the Schema Registry. If the schema version is valid, then its version ID and definition will be cached locally on the producer.

You can adjust the default cache period (24 hours) within the optional producer properties in step #3 of Installing SerDe Libraries.

3. Serialize and deliver records: If the record complies with the schema, the serializer decorates each record with the schema version ID, serializes the record based on the data format selected (AVRO, JSON, Protobuf, or other formats coming soon), compresses the record (optional producer configuration), and delivers it to the destination.
4. Consumers deserialize the data: Consumers reading this data use the Schema Registry deserializer library that parses the schema version ID from the record payload.
5. Deserializer may request the schema from the Schema Registry: If this is the first time the deserializer has seen records with a particular schema version ID, using the schema version ID the deserializer will request the schema from the Schema Registry and cache the schema locally on the consumer. If the Schema Registry cannot deserialize the record, the consumer can log the data from the record and move on, or halt the application.
6. The deserializer uses the schema to deserialize the record: When the deserializer retrieves the schema version ID from the Schema Registry, the deserializer decompresses the record (if record sent by producer is compressed) and uses the schema to deserialize the record. The application now processes the record.

Note

Encryption: Your clients communicate with the Schema Registry via API calls which encrypt data in-transit using TLS encryption over HTTPS. Schemas stored in the Schema Registry are always encrypted at rest using a service-managed AWS Key Management Service (AWS KMS) key.

Note

User Authorization: The Schema Registry supports identity-based IAM policies.

Getting started with schema registry

The following sections provide an overview and walk you through setting up and using Schema Registry. For information about schema registry concepts and components, see AWS Glue Schema registry.

Topics

- Installing SerDe Libraries
- Creating a registry
- Creating a schema
- Updating a schema or registry
- Deleting a schema or registry
- IAM examples for serializers
- IAM examples for deserializers
- Private connectivity using AWS PrivateLink
- Accessing Amazon CloudWatch metrics
- Sample AWS CloudFormation template for schema registry

Installing SerDe Libraries

> **Note**
>
> Prerequisites: Before completing the following steps, you will need to have a Amazon Managed Streaming for Apache Kafka (Amazon MSK) or Apache Kafka cluster running. Your producers and consumers need to be running on Java 8 or above.

The SerDe libraries provide a framework for serializing and deserializing data.

You will install the open source serializer for your applications producing data (collectively the "serializers"). The serializer handles serialization, compression, and the interaction with the Schema Registry. The serializer automatically extracts the schema from a record being written to a Schema Registry compatible destination, such as Amazon MSK. Likewise, you will install the open source deserializer on your applications consuming data.

To install the libraries on producers and consumers:

1. Inside both the producers’ and consumers’ pom.xml files, add this dependency via the code below:

   ```
   <dependency>
       <groupId>software.amazon.glue</groupId>
       <artifactId>schema-registry-serde</artifactId>
       <version>1.1.5</version>
   </dependency>
   ```

   Alternatively, you can clone the [AWS Glue Schema Registry Github repository]().

2. Setup your producers with these required properties:

   ```
   props.put(ProducerConfig.KEY_SERIALIZER_CLASS_CONFIG,
    StringSerializer.class.getName()); // Can replace StringSerializer.class.getName())
    with any other key serializer that you may use
   props.put(ProducerConfig.VALUE_SERIALIZER_CLASS_CONFIG,
    GlueSchemaRegistryKafkaSerializer.class.getName());
   props.put(AWSSchemaRegistryConstants.AWS_REGION, "us-east-2");
   properties.put(AWSSchemaRegistryConstants.DATA_FORMAT, "JSON"); // OR "AVRO"
   ```

   If there are no existing schemas, then auto-registration needs to be turned on (next step). If you do have a schema that you would like to apply, then replace "my-schema" with your schema name. Also the "registry-name" has to be provided if schema auto-registration is off. If the schema is created under the "default-registry" then registry name can be omitted.

3. (Optional) Set any of these optional producer properties. For detailed property descriptions, see [the ReadMe file]().

   ```
   props.put(AWSSchemaRegistryConstants.SCHEMA_AUTO_REGISTRATION_SETTING, "true"); // If
    not passed, uses "false"
   props.put(AWSSchemaRegistryConstants.SCHEMA_NAME, "my-schema"); // If not passed,
    uses transport name (topic name in case of Kafka, or stream name in case of Kinesis
    Data Streams)
   props.put(AWSSchemaRegistryConstants.REGISTRY_NAME, "my-registry"); // If not passed,
    uses "default-registry"
   props.put(AWSSchemaRegistryConstants.CACHE_TIME_TO_LIVE_MILLIS, "86400000"); // If
    not passed, uses 86400000 (24 Hours)
   props.put(AWSSchemaRegistryConstants.CACHE_SIZE, "10"); // default value is 200
   props.put(AWSSchemaRegistryConstants.COMPATIBILITY_SETTING, Compatibility.FULL); //
    Pass a compatibility mode. If not passed, uses Compatibility.BACKWARD
   ```

```
props.put(AWSSchemaRegistryConstants.DESCRIPTION, "This registry is used for several
 purposes."); // If not passed, constructs a description
props.put(AWSSchemaRegistryConstants.COMPRESSION_TYPE,
 AWSSchemaRegistryConstants.COMPRESSION.ZLIB); // If not passed, records are sent
 uncompressed
```

Auto-registration registers the schema version under the default registry ("default-registry"). If a `SCHEMA_NAME` is not specified in the previous step, then the topic name is inferred as `SCHEMA_NAME`.

See Schema versioning and compatibility for more information on compatibility modes.

4. Setup your consumers with these required properties:

```
props.put(ConsumerConfig.KEY_DESERIALIZER_CLASS_CONFIG,
 StringDeserializer.class.getName());
props.put(ConsumerConfig.VALUE_DESERIALIZER_CLASS_CONFIG,
 GlueSchemaRegistryKafkaDeserializer.class.getName());
props.put(AWSSchemaRegistryConstants.AWS_REGION, "us-east-2"); // Pass an AWS Region
props.put(AWSSchemaRegistryConstants.AVRO_RECORD_TYPE,
 AvroRecordType.GENERIC_RECORD.getName()); // Only required for AVRO data format
```

5. (Optional) Set these optional consumer properties. For detailed property descriptions, see the ReadMe file.

```
properties.put(AWSSchemaRegistryConstants.CACHE_TIME_TO_LIVE_MILLIS, "86400000"); //
 If not passed, uses 86400000
props.put(AWSSchemaRegistryConstants.CACHE_SIZE, "10"); // default value is 200
props.put(AWSSchemaRegistryConstants.SECONDARY_DESERIALIZER,
 "com.amazonaws.services.schemaregistry.deserializers.external.ThirdPartyDeserializer"); //
 For migration fall back scenario
```

Creating a registry

You may use the default registry or create as many new registries as necessary using the AWS Glue APIs or AWS Glue console.

AWS Glue APIs

You can use these steps to perform this task using the AWS Glue APIs.

To use the AWS CLI for the AWS Glue Schema Registry APIs, make sure to update your AWS CLI to the latest version.

To add a new registry, use the CreateRegistry action (Python: create_registry) API. Specify `RegistryName` as the name of the registry to be created, with a max length of 255, containing only letters, numbers, hyphens, underscores, dollar signs, or hash marks.

Specify a `Description` as a string not more than 2048 bytes long, matching the URI address multi-line string pattern.

Optionally, specify one or more `Tags` for your registry, as a map array of key-value pairs.

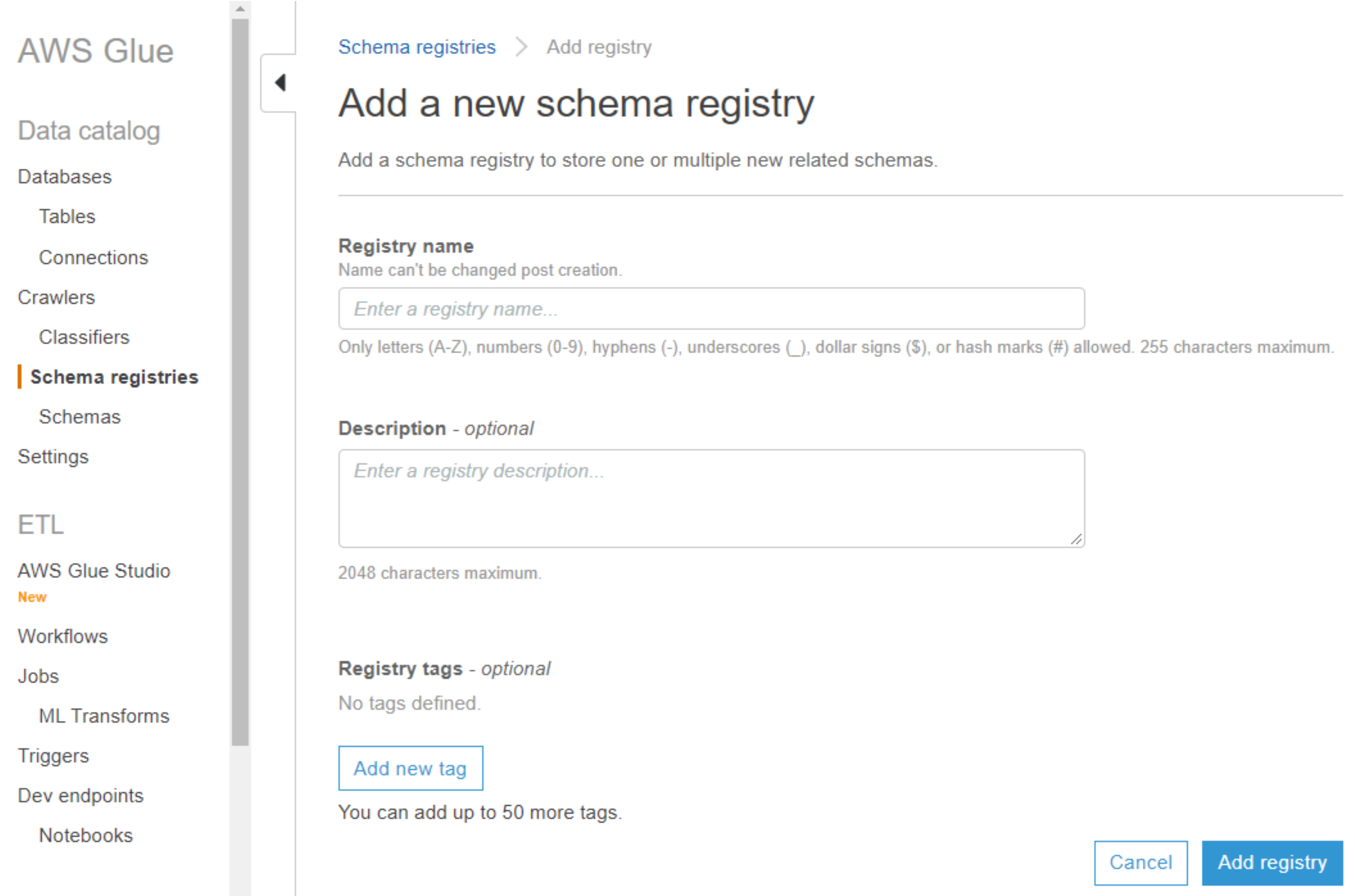
```
aws glue create-registry --registry-name registryName1 --description description
```

When your registry is created it is assigned an Amazon Resource Name (ARN), which you can view in the `RegistryArn` of the API response. Now that you've created a registry, create one or more schemas for that registry.

AWS Glue console

To add a new registry in the AWS Glue console:

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, under **Data catalog**, choose **Schema registries**.
3. Choose **Add registry**.
4. Enter a **Registry name** for the registry, consisting of letters, numbers, hyphens, or underscores. This name cannot be changed.
5. Enter a **Description** (optional) for the registry.
6. Optionally, apply one or more tags to your registry. Choose **Add new tag** and specify a **Tag key** and optionally a **Tag value**.
7. Choose **Add registry**.


AWS Glue
Data catalog
Databases
Tables
Connections
Crawlers
Classifiers
Schema registries
Schemas
Settings
ETL
AWS Glue Studio
New
Workflows
Jobs
ML Transforms
Triggers
Dev endpoints
Notebooks
Schema registries
Add registry

Add a new schema registry

Add a schema registry to store one or multiple new related schemas.

Registry name
Name can't be changed post creation.
Enter a registry name...
Only letters (A-Z), numbers (0-9), hyphens (-), underscores (_), dollar signs ($), or hash marks (#) allowed. 255 characters maximum.

Description - optional
Enter a registry description...
2048 characters maximum.

Registry tags - optional
No tags defined.

You can add up to 50 more tags.

Cancel
Add registry

When your registry is created it is assigned an Amazon Resource Name (ARN), which you can view by choosing the registry from the list in **Schema registries**. Now that you've created a registry, create one or more schemas for that registry.

Dealing with a specific record (JAVA POJO) for JSON

You can use a plain old Java object (POJO) and pass the object as a record. This is similar to the notion of a specific record in AVRO. The mbknor-jackson-jsonschema can generate a JSON schema for the POJO passed. This library can also inject additional information in the JSON schema.

The AWS Glue Schema Registry library uses the injected "className" field in schema to provide a fully classified class name. The "className" field is used by the deserializer to deserialize into an object of that class.

```
 Example class :

@JsonSchemaDescription("This is a car")
@JsonSchemaTitle("Simple Car Schema")
```

```
@Builder
@AllArgsConstructor
@EqualsAndHashCode
// Fully qualified class name to be added to an additionally injected property
// called className for deserializer to determine which class to deserialize
// the bytes into
@JsonSchemaInject(
        strings = {@JsonSchemaString(path = "className",
                value =
 "com.amazonaws.services.schemaregistry.integrationtests.generators.Car")}
)
// List of annotations to help infer JSON Schema are defined by https://github.com/
mbknor/mbknor-jackson-jsonSchema
public class Car {
    @JsonProperty(required = true)
    private String make;

    @JsonProperty(required = true)
    private String model;

    @JsonSchemaDefault("true")
    @JsonProperty
    public boolean used;

    @JsonSchemaInject(ints = {@JsonSchemaInt(path = "multipleOf", value = 1000)})
    @Max(200000)
    @JsonProperty
    private int miles;

    @Min(2000)
    @JsonProperty
    private int year;

    @JsonProperty
    private Date purchaseDate;

    @JsonProperty
    @JsonFormat(shape = JsonFormat.Shape.NUMBER)
    private Date listedDate;

    @JsonProperty
    private String[] owners;

    @JsonProperty
    private Collection<Float> serviceChecks;

    // Empty constructor is required by Jackson to deserialize bytes
    // into an Object of this class
    public Car() {}
}
```

Creating a schema

You can create a schema using the AWS Glue APIs or the AWS Glue console.

AWS Glue APIs

You can use these steps to perform this task using the AWS Glue APIs.

To add a new schema, use the CreateSchema action (Python: create_schema) API.

Specify a `RegistryId` structure to indicate a registry for the schema. Or, omit the `RegistryId` to use the default registry.

Specify a `SchemaName` consisting of letters, numbers, hyphens, or underscores, and `DataFormat` as **`AVRO`** or **`JSON`**. `DataFormat` once set on a schema is not changeable.

Specify a `Compatibility` mode:

- *Backward (recommended)* — Consumer can read both current and previous version.
- *Backward all* — Consumer can read current and all previous versions.
- *Forward* — Consumer can read both current and subsequent version.
- *Forward all* — Consumer can read both current and all subsequent versions.
- *Full* — Combination of Backward and Forward.
- *Full all* — Combination of Backward all and Forward all.
- *None* — No compatibility checks are performed.
- *Disabled* — Prevent any versioning for this schema.

Optionally, specify `Tags` for your schema.

Specify a `SchemaDefinition` to define the schema in Avro, JSON, or Protobuf data format. See the examples.

For Avro data format:

```
aws glue create-schema --registry-id RegistryName="registryName1" --schema-name
 testschema --compatibility NONE --data-format AVRO --schema-definition "{\"type\":
 \"record\", \"name\": \"r1\", \"fields\": [ {\"name\": \"f1\", \"type\": \"int\"},
 {\"name\": \"f2\", \"type\": \"string\"} ]}"
```

```
aws glue create-schema --registry-id RegistryArn="arn:aws:glue:us-
east-2:901234567890:registry/registryName1" --schema-name testschema --compatibility
 NONE --data-format AVRO  --schema-definition "{\"type\": \"record\", \"name\": \"r1\",
 \"fields\": [ {\"name\": \"f1\", \"type\": \"int\"}, {\"name\": \"f2\", \"type\":
 \"string\"} ]}"
```

For JSON data format:

```
aws glue create-schema --registry-id RegistryName="registryName" --schema-name
 testSchemaJson --compatibility NONE --data-format JSON --schema-definition "{\"$schema
\": \"http://json-schema.org/draft-07/schema#\",\"type\":\"object\",\"properties\":
{\"f1\":{\"type\":\"string\"}}}"
```

```
aws glue create-schema --registry-id RegistryArn="arn:aws:glue:us-
east-2:901234567890:registry/registryName" --schema-name testSchemaJson --compatibility
 NONE --data-format JSON --schema-definition "{\"$schema\": \"http://json-schema.org/
draft-07/schema#\",\"type\":\"object\",\"properties\":{\"f1\":{\"type\":\"string\"}}}"
```

For Protobuf data format:

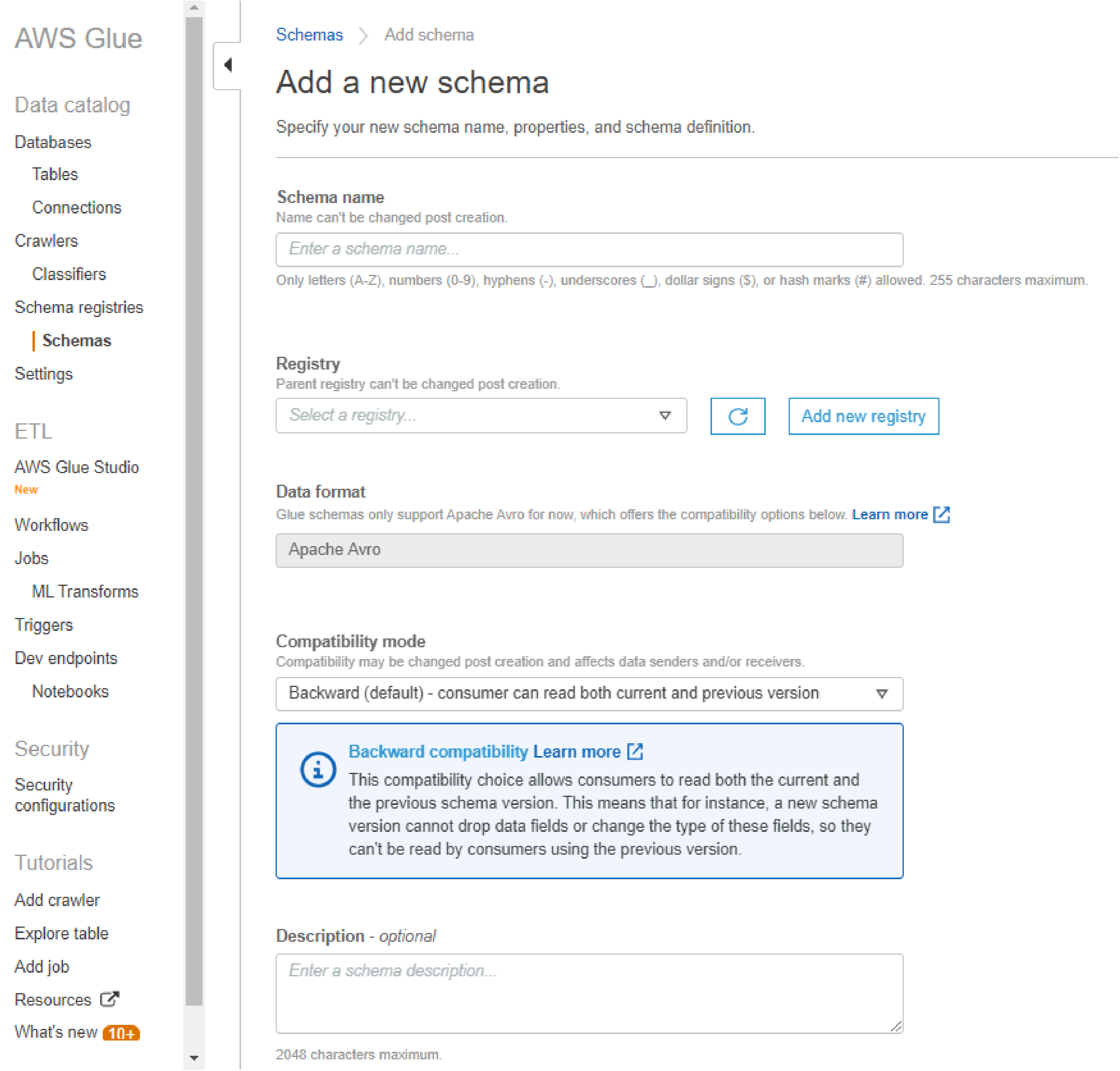
```
aws glue create-schema --registry-id RegistryName="registryName" --schema-name
 testSchemaProtobuf --compatibility NONE --data-format PROTOBUF --schema-definition
 "syntax = \"proto2\";package org.test;message Basic { optional int32 basic = 1;}"
```

```
aws glue create-schema --registry-id RegistryArn="arn:aws:glue:us-
east-2:901234567890:registry/registryName" --schema-name testSchemaProtobuf
 --compatibility NONE --data-format PROTOBUF --schema-definition "syntax =
 \"proto2\";package org.test;message Basic { optional int32 basic = 1;}"
```

AWS Glue console

To add a new schema using the AWS Glue console:

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, under **Data catalog**, choose **Schemas**.
3. Choose **Add schema**.
4. Enter a **Schema name**, consisting of letters, numbers, hyphens, underscores, dollar signs, or hashmarks. This name cannot be changed.
5. Choose the **Registry** where the schema will be stored from the drop-down menu. The parent registry cannot be changed post-creation.
6. Leave the **Data format** as *Apache Avro* or *JSON*. This format applies to all versions of this schema.
7. Choose a **Compatibility mode**.
    - *Backward (recommended)* — receiver can read both current and previous versions.
    - *Backward All* — receiver can read current and all previous versions.
    - *Forward* — sender can write both current and previous versions.
    - *Forward All* — sender can write current and all previous versions.
    - *Full* — combination of Backward and Forward.
    - *Full All* — combination of Backward All and Forward All.
    - *None* — no compatibility checks performed.
    - *Disabled* — prevent any versioning for this schema.
8. Enter an optional **Description** for the registry of up to 250 characters.


AWS Glue
Data catalog
Databases
Tables
Connections
Crawlers
Classifiers
Schema registries
Schemas
Settings
ETL
AWS Glue Studio
New
Workflows
Jobs
ML Transforms
Triggers
Dev endpoints
Notebooks
Security
Security configurations
Tutorials
Add crawler
Explore table
Add job
Resources
What's new 10+
Schemas > Add schema
Add new registry

Add a new schema

Specify your new schema name, properties, and schema definition.

Schema name
Name can't be changed post creation.
Enter a schema name...
Only letters (A-Z), numbers (0-9), hyphens (-), underscores (_), dollar signs ($), or hash marks (#) allowed. 255 characters maximum.

Registry
Parent registry can't be changed post creation.
Select a registry...

Data format
Glue schemas only support Apache Avro for now, which offers the compatibility options below. Learn more
Apache Avro

Compatibility mode
Compatibility may be changed post creation and affects data senders and/or receivers.
Backward (default) - consumer can read both current and previous version

Backward compatibility Learn more
This compatibility choice allows consumers to read both the current and the previous schema version. This means that for instance, a new schema version cannot drop data fields or change the type of these fields, so they can't be read by consumers using the previous version.

Description - *optional*
Enter a schema description...
2048 characters maximum.

9. Optionally, apply one or more tags to your schema. Choose **Add new tag** and specify a **Tag key** and optionally a **Tag value**.

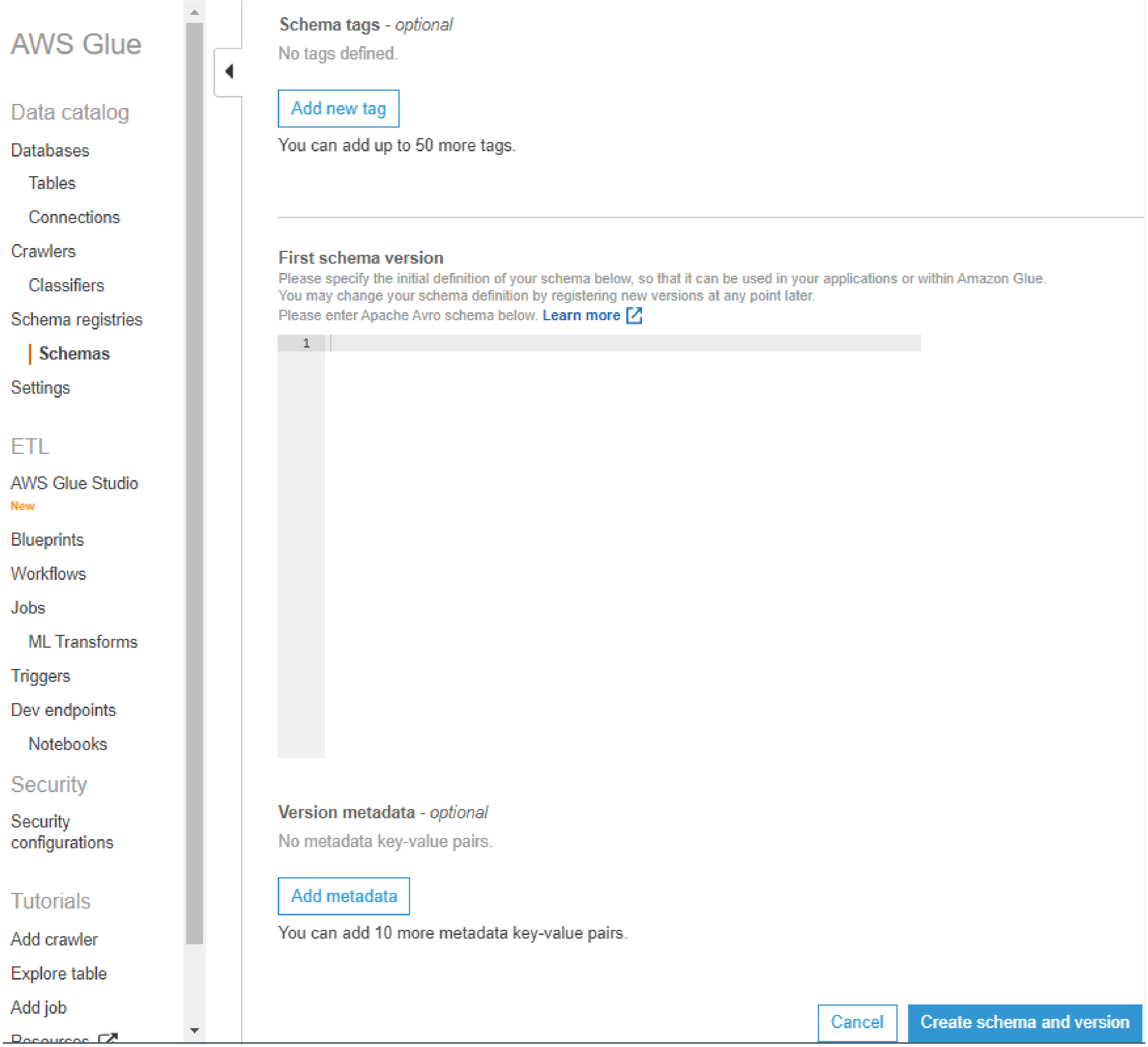
10In the **First schema version** box, enter or paste your initial schema. .

   For Avro format, see Working with Avro data format

   For JSON format, see Working with JSON data format

11. Optionally, choose **Add metadata** to add version metadata to annotate or classify your schema version.

12. Choose **Create schema and version**.


AWS Glue
Data catalog
Databases
Tables
Connections
Crawlers
Classifiers
Schema registries
Schemas
Settings
ETL
AWS Glue Studio
New
Blueprints
Workflows
Jobs
ML Transforms
Triggers
Dev endpoints
Notebooks
Security
Security configurations
Tutorials
Add crawler
Explore table
Add job
Add new tag
Add metadata
Cancel
Create schema and version

Schema tags - *optional*

No tags defined.

You can add up to 50 more tags.

First schema version

Please specify the initial definition of your schema below, so that it can be used in your applications or within Amazon Glue. You may change your schema definition by registering new versions at any point later.

Please enter Apache Avro schema below. Learn more

Version metadata - *optional*

No metadata key-value pairs.

You can add 10 more metadata key-value pairs.

The schema is created and appears in the list under **Schemas**.

Working with Avro data format

Avro provides data serialization and data exchange services. Avro stores the data definition in JSON format making it easy to read and interpret. The data itself is stored in binary format.

For information on defining an Apache Avro schema, see the Apache Avro specification.

Working with JSON data format

Data can be serialized with JSON format. JSON Schema format defines the standard for JSON Schema format.

Updating a schema or registry

Once created you can edit your schemas, schema versions, or registry.

Updating a registry

You can update a registry using the AWS Glue APIs or the AWS Glue console. The name of an existing registry cannot be edited. You can edit the description for a registry.

AWS Glue APIs

To update an existing registry, use the UpdateRegistry action (Python: update_registry) API.

Specify a `RegistryId` structure to indicate the registry that you want to update. Pass a `Description` to change the description for a registry.

```
aws glue update-registry --description updatedDescription --registry-id
 RegistryArn="arn:aws:glue:us-east-2:901234567890:registry/registryName1"
```

AWS Glue console

To update a registry using the AWS Glue console:

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, under **Data catalog**, choose **Schema registries**.
3. Choose a registry from the the list of registries, by checking its box.
4. In the **Action** menu, choose **Edit registry**.

Updating a schema

You can update the description or compatibility setting for a schema.

To update an existing schema, use the UpdateSchema action (Python: update_schema) API.

Specify a `SchemaId` structure to indicate the schema that you want to update. One of `VersionNumber` or `Compatibility` has to be provided.

Code example 11:

```
aws glue update-schema --description testDescription --schema-id
 SchemaName="testSchema1",RegistryName="registryName1" --schema-version-number
 LatestVersion=true --compatibility NONE
```

```
aws glue update-schema --description testDescription --schema-id
 SchemaArn="arn:aws:glue:us-east-2:901234567890:schema/registryName1/testSchema1" --
schema-version-number LatestVersion=true --compatibility NONE
```

Adding a schema version

When you add a schema version, you will need to compare the versions to make sure the new schema will be accepted.

To add a new version to an existing schema, use the RegisterSchemaVersion action (Python: register_schema_version) API.

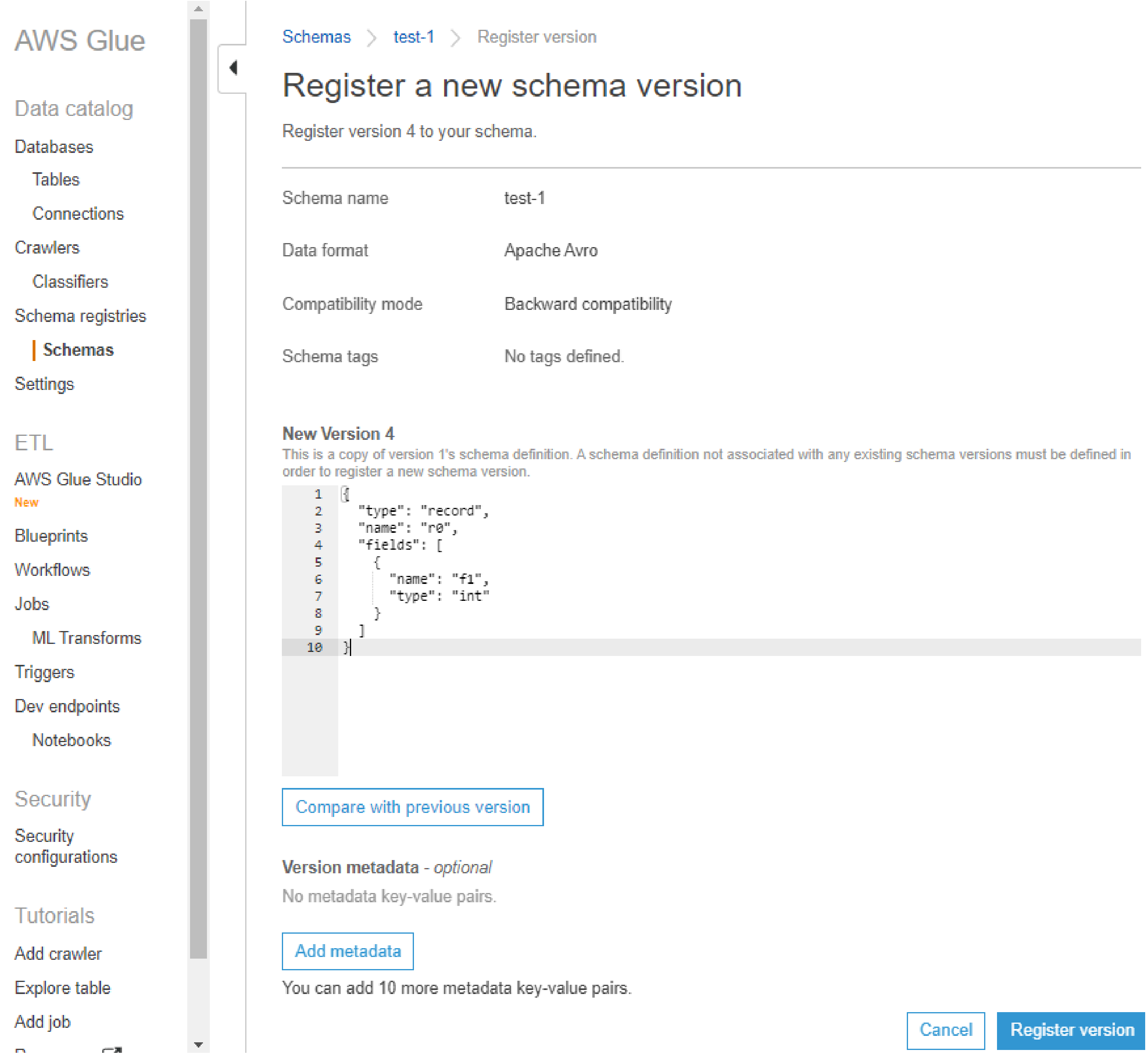
Specify a `SchemaId` structure to indicate the schema for which you want to add a version, and a `SchemaDefinition` to define the schema.

Code example 12:

```
aws glue register-schema-version --schema-definition "{\"type\": \"record\", \"name\":
 \"r1\", \"fields\": [ {\"name\": \"f1\", \"type\": \"int\"}, {\"name\": \"f2\", \"type
\": \"string\"} ]}" --schema-id SchemaArn="arn:aws:glue:us-east-1:901234567890:schema/
registryName/testschema"
```

```
aws glue register-schema-version --schema-definition "{\"type\": \"record\", \"name\":
 \"r1\", \"fields\": [ {\"name\": \"f1\", \"type\": \"int\"}, {\"name\": \"f2\", \"type
\": \"string\"} ]}" --schema-id SchemaName="testschema",RegistryName="testregistry"
```

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, under **Data catalog**, choose **Schemas**.
3. Choose the schema from the the list of schemas, by checking its box.
4. Choose one or more schemas from the list, by checking the boxes.
5. In the **Action** menu, choose **Register new version**.
6. In the **New version** box, enter or paste your new schema.
7. Choose **Compare with previous version** to see differences with the previous schema version.
8. Optionally, choose **Add metadata** to add version metadata to annotate or classify your schema version. Enter **Key** and optional **Value**.
9. Choose **Register version**.


AWS Glue
Data catalog
Databases
Tables
Connections
Crawlers
Classifiers
Schema registries
Schemas
Settings
ETL
AWS Glue Studio
New
Blueprints
Workflows
Jobs
ML Transforms
Triggers
Dev endpoints
Notebooks
Security
Security configurations
Tutorials
Add crawler
Explore table
Add job
Schemas > test-1 > Register version
Compare with previous version
Add metadata
Cancel
Register version

Register a new schema version

Register version 4 to your schema.

Schema name: test-1

Data format: Apache Avro

Compatibility mode: Backward compatibility

Schema tags: No tags defined.

New Version 4

This is a copy of version 1's schema definition. A schema definition not associated with any existing schema versions must be defined in order to register a new schema version.

```
{
  "type": "record",
  "name": "r0",
  "fields": [
    {
      "name": "f1",
      "type": "int"
    }
  ]
}
```

Version metadata - *optional*

No metadata key-value pairs.

You can add 10 more metadata key-value pairs.

The schema(s) version appears in the list of versions. If the version changed the compatibility mode, the version will be marked as a checkpoint.

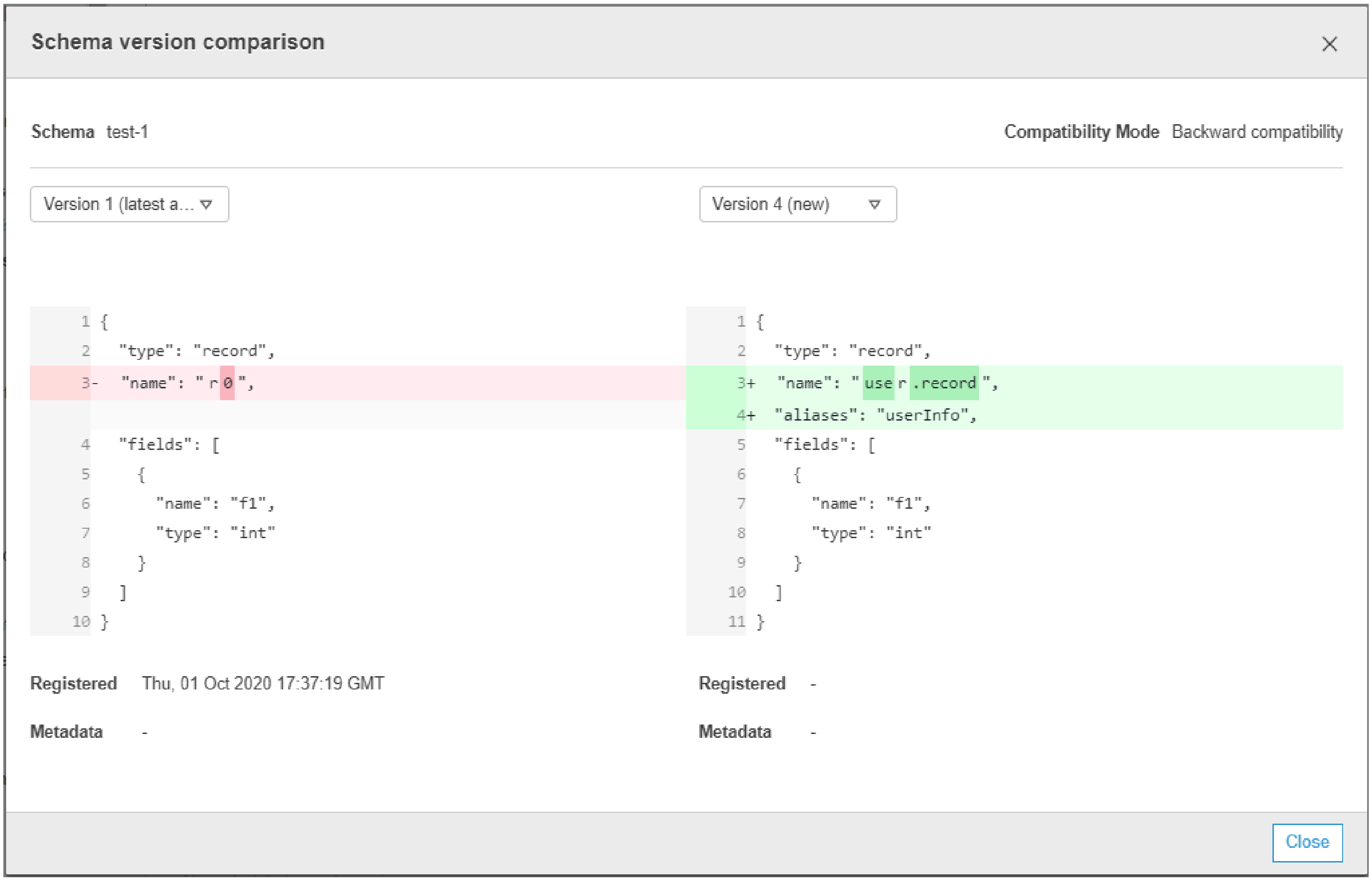
Example of a schema version comparison

When you choose to **Compare with previous version**, you will see the previous and new versions displayed together. Changed information will be highlighted as follows:

- *Yellow*: indicates changed information.

- *Green*: indicates content added in the latest version.
- *Red*: indicates content removed in the latest version.

You can also compare against earlier versions.



Schema test-1

Compatibility Mode Backward compatibility

```
{
  "type": "record",
  "name": "r0",

  "fields": [
    {
      "name": "f1",
      "type": "int"
    }
  ]
}
```

```
{
  "type": "record",
  "name": "user.record",
  "aliases": "userInfo",
  "fields": [
    {
      "name": "f1",
      "type": "int"
    }
  ]
}
```

Registered Thu, 01 Oct 2020 17:37:19 GMT

Metadata -

Registered -

Metadata -

Deleting a schema or registry

Deleting a schema, a schema version, or a registry are permanent actions that cannot be undone.

Deleting a schema

You may want to delete a schema when it will no longer be used within a registry, using the AWS Management Console, or the DeleteSchema action (Python: delete_schema) API.

Deleting one or more schemas is a permanent action that cannot be undone. Make sure that the schema or schemas are no longer needed.

To delete a schema from the registry, call the DeleteSchema action (Python: delete_schema) API, specifying the `SchemaId` structure to identify the schema.

For example:

```
aws glue delete-schema --schema-id SchemaArn="arn:aws:glue:us-
east-2:901234567890:schema/registryName1/schemaname"
```

```
aws glue delete-schema --schema-id SchemaName="TestSchema6-
deleteschemabyname",RegistryName="default-registry"
```

AWS Glue console

To delete a schema from the AWS Glue console:

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, under **Data catalog**, choose **Schema registries**.
3. Choose the registry that contains your schema from the the list of registries.
4. Choose one or more schemas from the list, by checking the boxes.
5. In the **Action** menu, choose **Delete schema**.
6. Enter the text **`Delete`** in the field to confirm deletion.
7. Choose **Delete**.

The schema(s) you specified are deleted from the registry.

Deleting a schema version

As schemas accumulate in the registry, you may want to delete unwanted schema versions using the AWS Management Console, or the DeleteSchemaVersions action (Python: delete_schema_versions) API. Deleting one or more schema versions is a permanent action that cannot be undone. Make sure that the schema versions are no longer needed.

When deleting schema versions, take note of the following constraints:

- You cannot delete a check-pointed version.
- The range of contiguous versions cannot be more than 25.
- The latest schema version must not be in a pending state.

Specify the `SchemaId` structure to identify the schema, and specify `Versions` as a range of versions to delete. For more information on specifying a version or range of versions, see DeleteRegistry action (Python: delete_registry). The schema versions you specified are deleted from the registry.

Calling the ListSchemaVersions action (Python: list_schema_versions) API after this call will list the status of the deleted versions.

For example:

```
aws glue delete-schema-versions --schema-id
 SchemaName="TestSchema6",RegistryName="default-registry" --versions "1-1"
```

```
aws glue delete-schema-versions --schema-id SchemaArn="arn:aws:glue:us-
east-2:901234567890:schema/default-registry/TestSchema6-NON-Existent" --versions "1-1"
```

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, under **Data catalog**, choose **Schema registries**.
3. Choose the registry that contains your schema from the the list of registries.
4. Choose one or more schemas from the list, by checking the boxes.
5. In the **Action** menu, choose **Delete schema**.
6. Enter the text **`Delete`** in the field to confirm deletion.
7. Choose **Delete**.

The schema versions you specified are deleted from the registry.

Deleting a registry

You may want to delete a registry when the schemas it contains should no longer be organized under that registry. You will need to reassign those schemas to another registry.

Deleting one or more registries is a permanent action that cannot be undone. Make sure that the registry or registries no longer needed.

The default registry can be deleted using the AWS CLI.

AWS Glue API

To delete the entire registry including the schema and all of its versions, call the DeleteRegistry action (Python: delete_registry) API. Specify a `RegistryId` structure to identify the registry.

For example:

```
aws glue delete-registry --registry-id RegistryArn="arn:aws:glue:us-
east-2:901234567890:registry/registryName1"
```

```
aws glue delete-registry --registry-id RegistryName="TestRegistry-deletebyname"
```

To get the status of the delete operation, you can call the `GetRegistry` API after the asynchronous call.

AWS Glue console

To delete a registry from the AWS Glue console:

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, under **Data catalog**, choose **Schema registries**.
3. Choose a registry from the list, by checking a box.
4. In the **Action** menu, choose **Delete registry**.
5. Enter the text **`Delete`** in the field to confirm deletion.
6. Choose **Delete**.

The registries you selected are deleted from AWS Glue.

IAM examples for serializers

> **Note**
>
> AWS managed policies grant necessary permissions for common use cases. For information on using managed policies to manage the schema registry, see AWS managed (predefined) policies for AWS Glue.

For serializers, you should create a minimal policy similar to that below to give you the ability to find the `schemaVersionId` for a given schema definition. Note, you should have read permissions on the registry in order to read the schemas in the registry. You can limit the registries that can be read by using the `Resource` clause.

Code example 13:

```
{
    "Sid" : "GetSchemaByDefinition",
    "Effect" : "Allow",
    "Action" :
 [
        "glue:GetSchemaByDefinition"
    ],
        "Resource" : ["arn:aws:glue:us-east-2:012345678:registry/registryname-1",
                      "arn:aws:glue:us-east-2:012345678:schema/registryname-1/
schemaname-1",
                      "arn:aws:glue:us-east-2:012345678:schema/registryname-1/
schemaname-2"
                     ]
}
```

Further, you can also allow producers to create new schemas and versions by including the following extra methods. Note, you should be able to inspect the registry in order to add/remove/evolve the schemas inside it. You can limit the registries that can be inspected by using the `Resource` clause.

Code example 14:

```
{
    "Sid" : "RegisterSchemaWithMetadata",
    "Effect" : "Allow",
    "Action" :
 [
        "glue:GetSchemaByDefinition",
        "glue:CreateSchema",
        "glue:RegisterSchemaVersion",
        "glue:PutSchemaVersionMetadata",
    ],
    "Resource" : ["arn:aws:glue:aws-region:123456789012:registry/registryname-1",
                  "arn:aws:glue:aws-region:123456789012:schema/registryname-1/
schemaname-1",
                  "arn:aws:glue:aws-region:123456789012:schema/registryname-1/
schemaname-2"
                 ]
}
```

IAM examples for deserializers

For deserializers (consumer side), you should create a policy similar to that below to allow the deserializer to fetch the schema from the Schema Registry for deserialization. Note, you should be able to inspect the registry in order to fetch the schemas inside it.

Code example 15:

```
{
    "Sid" : "GetSchemaVersion",
    "Effect" : "Allow",
    "Action" :
 [
        "glue:GetSchemaVersion"
    ],
    "Resource" : ["*"]
}
```

Private connectivity using AWS PrivateLink

You can use AWS PrivateLink to connect your data producer's VPC to AWS Glue by defining an interface VPC endpoint for AWS Glue. When you use a VPC interface endpoint, communication between your VPC and AWS Glue is conducted entirely within the AWS network. For more information, see Using AWS Glue with VPC Endpoints.

Accessing Amazon CloudWatch metrics

Amazon CloudWatch metrics are available as part of CloudWatch's free tier. You can access these metrics in the CloudWatch console. API-Level metrics include CreateSchema (Success and Latency), GetSchemaByDefinition, (Success and Latency), GetSchemaVersion (Success and Latency), RegisterSchemaVersion (Success and Latency), PutSchemaVersionMetadata (Success and Latency). Resource-level metrics include Registry.ThrottledByLimit, SchemaVersion.ThrottledByLimit, SchemaVersion.Size.

Sample AWS CloudFormation template for schema registry

The following is a sample template for creating Schema Registry resources in AWS CloudFormation. To create this stack in your account, copy the above template into a file `SampleTemplate.yaml`, and run the following command:

```
aws cloudformation create-stack --stack-name ABCSchemaRegistryStack --template-body
 "'cat SampleTemplate.yaml'"
```

This example uses `AWS::Glue::Registry` to create a registry, `AWS::Glue::Schema` to create a schema, `AWS::Glue::SchemaVersion` to create a schema version, and `AWS::Glue::SchemaVersionMetadata` to populate schema version metadata.

```
Description: "A sample CloudFormation template for creating Schema Registry resources."
Resources:
  ABCRegistry:
    Type: "AWS::Glue::Registry"
    Properties:
      Name: "ABCSchemaRegistry"
      Description: "ABC Corp. Schema Registry"
      Tags:
        - Key: "Project"
          Value: "Foo"
  ABCSchema:
    Type: "AWS::Glue::Schema"
    Properties:
      Registry:
        Arn: !Ref ABCRegistry
      Name: "TestSchema"
      Compatibility: "NONE"
      DataFormat: "AVRO"
      SchemaDefinition: >
        {"namespace":"foo.avro","type":"record","name":"user","fields":
[{"name":"name","type":"string"},{"name":"favorite_number","type":"int"}]}
      Tags:
        - Key: "Project"
          Value: "Foo"
  SecondSchemaVersion:
    Type: "AWS::Glue::SchemaVersion"
    Properties:
      Schema:
        SchemaArn: !Ref ABCSchema
      SchemaDefinition: >
        {"namespace":"foo.avro","type":"record","name":"user","fields":
[{"name":"status","type":"string", "default":"ON"}, {"name":"name","type":"string"},
{"name":"favorite_number","type":"int"}]}
  FirstSchemaVersionMetadata:
    Type: "AWS::Glue::SchemaVersionMetadata"
```

```
      Properties:
        SchemaVersionId: !GetAtt ABCSchema.InitialSchemaVersionId
        Key: "Application"
        Value: "Kinesis"
    SecondSchemaVersionMetadata:
      Type: "AWS::Glue::SchemaVersionMetadata"
      Properties:
        SchemaVersionId: !Ref SecondSchemaVersion
        Key: "Application"
        Value: "Kinesis"
```

Integrating with AWS Glue Schema Registry

These sections describe integrations with AWS Glue schema registry. The examples in these section show a schema with AVRO data format. For more examples, including schemas with JSON data format, see the integration tests and ReadMe information in the AWS Glue Schema Registry open source repository.

Topics

- Use case: Connecting Schema Registry to Amazon MSK or Apache Kafka
- Use case: Integrating Amazon Kinesis Data Streams with the AWS Glue Schema Registry
- Use case: Amazon Managed Service for Apache Flink
- Use Case: Integration with AWS Lambda
- Use case: AWS Glue Data Catalog
- Use case: AWS Glue streaming
- Use case: Apache Kafka Streams

Use case: Connecting Schema Registry to Amazon MSK or Apache Kafka

Let's assume you are writing data to an Apache Kafka topic, and you can follow these steps to get started.

1. Create an Amazon Managed Streaming for Apache Kafka (Amazon MSK) or Apache Kafka cluster with at least one topic. If creating an Amazon MSK cluster, you can use the AWS Management Console. Follow these instructions: Getting Started Using Amazon MSK in the *Amazon Managed Streaming for Apache Kafka Developer Guide*.

2. Follow the Installing SerDe Libraries step above.
3. To create schema registries, schemas, or schema versions, follow the instructions under the Getting started with schema registry section of this document.
4. Start your producers and consumers to use the Schema Registry to write and read records to/from the Amazon MSK or Apache Kafka topic. Example producer and consumer code can be found in the ReadMe file from the Serde libraries. The Schema Registry library on the producer will automatically serialize the record and decorate the record with a schema version ID.
5. If the schema of this record has been inputted, or if auto-registration is turned on, then the schema will have been registered in the Schema Registry.
6. The consumer reading from the Amazon MSK or Apache Kafka topic, using the AWS Glue Schema Registry library, will automatically lookup the schema from the Schema Registry.

Use case: Integrating Amazon Kinesis Data Streams with the AWS Glue Schema Registry

This integration requires that you have an existing Amazon Kinesis data stream. For more information, see Getting Started with Amazon Kinesis Data Streams in the *Amazon Kinesis Data Streams Developer Guide*.

There are two ways that you can interact with data in a Kinesis data stream.

- Through the Kinesis Producer Library (KPL) and Kinesis Client Library (KCL) libraries in Java. Multi-language support is not provided.
- Through the `PutRecords`, `PutRecord`, and `GetRecords` Kinesis Data Streams APIs available in the AWS SDK for Java.

If you currently use the KPL/KCL libraries, we recommend continuing to use that method. There are updated KCL and KPL versions with Schema Registry integrated, as shown in the examples. Otherwise, you can use the sample code to leverage the AWS Glue Schema Registry if using the KDS APIs directly.

Schema Registry integration is only available with KPL v0.14.2 or later and with KCL v2.3 or later. Schema Registry integration with JSON data format is available with KPL v0.14.8 or later and with KCL v2.3.6 or later.

Interacting with Data Using Kinesis SDK V2

This section describes interacting with Kinesis using Kinesis SDK V2

```
// Example JSON Record, you can construct a AVRO record also
private static final JsonDataWithSchema record =
 JsonDataWithSchema.builder(schemaString, payloadString);
private static final DataFormat dataFormat = DataFormat.JSON;

//Configurations for Schema Registry
GlueSchemaRegistryConfiguration gsrConfig = new GlueSchemaRegistryConfiguration("us-
east-1");

GlueSchemaRegistrySerializer glueSchemaRegistrySerializer =
        new GlueSchemaRegistrySerializerImpl(awsCredentialsProvider, gsrConfig);
GlueSchemaRegistryDataFormatSerializer dataFormatSerializer =
        new GlueSchemaRegistrySerializerFactory().getInstance(dataFormat, gsrConfig);

Schema gsrSchema =
        new Schema(dataFormatSerializer.getSchemaDefinition(record), dataFormat.name(),
 "MySchema");

byte[] serializedBytes = dataFormatSerializer.serialize(record);

byte[] gsrEncodedBytes = glueSchemaRegistrySerializer.encode(streamName, gsrSchema,
 serializedBytes);

PutRecordRequest putRecordRequest = PutRecordRequest.builder()
        .streamName(streamName)
        .partitionKey("partitionKey")
        .data(SdkBytes.fromByteArray(gsrEncodedBytes))
        .build();
shardId = kinesisClient.putRecord(putRecordRequest)
        .get()
        .shardId();

GlueSchemaRegistryDeserializer glueSchemaRegistryDeserializer = new
 GlueSchemaRegistryDeserializerImpl(awsCredentialsProvider, gsrConfig);

GlueSchemaRegistryDataFormatDeserializer gsrDataFormatDeserializer =
        glueSchemaRegistryDeserializerFactory.getInstance(dataFormat, gsrConfig);

GetShardIteratorRequest getShardIteratorRequest = GetShardIteratorRequest.builder()
        .streamName(streamName)
        .shardId(shardId)
        .shardIteratorType(ShardIteratorType.TRIM_HORIZON)
        .build();

String shardIterator = kinesisClient.getShardIterator(getShardIteratorRequest)
        .get()
        .shardIterator();

GetRecordsRequest getRecordRequest = GetRecordsRequest.builder()
        .shardIterator(shardIterator)
        .build();
GetRecordsResponse recordsResponse = kinesisClient.getRecords(getRecordRequest)
        .get();

List<Object> consumerRecords = new ArrayList<>();
List<Record> recordsFromKinesis = recordsResponse.records();

for (int i = 0; i < recordsFromKinesis.size(); i++) {
    byte[] consumedBytes = recordsFromKinesis.get(i)
            .data()
            .asByteArray();

    Schema gsrSchema = glueSchemaRegistryDeserializer.getSchema(consumedBytes);
    Object decodedRecord =
 gsrDataFormatDeserializer.deserialize(ByteBuffer.wrap(consumedBytes),

 gsrSchema.getSchemaDefinition());
    consumerRecords.add(decodedRecord);
}
```

Interacting with data using the KPL/KCL libraries

This section describes integrating Kinesis Data Streams with Schema Registry using the KPL/KCL libraries. For more information on using KPL/KCL, see Developing Producers Using the Amazon Kinesis Producer Library in the *Amazon Kinesis Data Streams Developer Guide*.

Setting up the Schema Registry in KPL

1. Define the schema definition for the data, data format and schema name authored in the AWS Glue Schema Registry.
2. Optionally configure the `GlueSchemaRegistryConfiguration` object.
3. Pass the schema object to the `addUserRecord API`.

```
private static final String SCHEMA_DEFINITION = "{"namespace": "example.avro",\n"
+ " "type": "record",\n"
+ " "name": "User",\n"
+ " "fields": [\n"
+ " {"name": "name", "type": "string"},\n"
+ " {"name": "favorite_number", "type": ["int", "null"]},\n"
+ " {"name": "favorite_color", "type": ["string", "null"]}\n"
+ " ]\n"
+ "}";

KinesisProducerConfiguration config = new KinesisProducerConfiguration();
config.setRegion("us-west-1")

//[Optional] configuration for Schema Registry.

GlueSchemaRegistryConfiguration schemaRegistryConfig =
new GlueSchemaRegistryConfiguration("us-west-1");

schemaRegistryConfig.setCompression(true);

config.setGlueSchemaRegistryConfiguration(schemaRegistryConfig);

///Optional configuration ends.

final KinesisProducer producer =
      new KinesisProducer(config);

final ByteBuffer data = getDataToSend();

com.amazonaws.services.schemaregistry.common.Schema gsrSchema =
    new Schema(SCHEMA_DEFINITION, DataFormat.AVRO.toString(), "demoSchema");

ListenableFuture<UserRecordResult> f = producer.addUserRecord(
config.getStreamName(), TIMESTAMP, Utils.randomExplicitHashKey(), data, gsrSchema);

private static ByteBuffer getDataToSend() {
      org.apache.avro.Schema avroSchema =
        new org.apache.avro.Schema.Parser().parse(SCHEMA_DEFINITION);

      GenericRecord user = new GenericData.Record(avroSchema);
      user.put("name", "Emily");
      user.put("favorite_number", 32);
      user.put("favorite_color", "green");
      ByteArrayOutputStream outBytes = new ByteArrayOutputStream();
      Encoder encoder = EncoderFactory.get().directBinaryEncoder(outBytes, null);
      new GenericDatumWriter<>(avroSchema).write(user, encoder);
      encoder.flush();
      return ByteBuffer.wrap(outBytes.toByteArray());
 }
```

Setting up the Kinesis client library

You will develop your Kinesis Client Library consumer in Java. For more information, see Developing a Kinesis Client Library Consumer in Java in the *Amazon Kinesis Data Streams Developer Guide*.

1. Create an instance of `GlueSchemaRegistryDeserializer` by passing a `GlueSchemaRegistryConfiguration` object.
2. Pass the `GlueSchemaRegistryDeserializer` to `retrievalConfig.glueSchemaRegistryDeserializer`.
3. Access the schema of incoming messages by calling `kinesisClientRecord.getSchema()`.

```
GlueSchemaRegistryConfiguration schemaRegistryConfig =
    new GlueSchemaRegistryConfiguration(this.region.toString());

 GlueSchemaRegistryDeserializer glueSchemaRegistryDeserializer =
    new
 GlueSchemaRegistryDeserializerImpl(DefaultCredentialsProvider.builder().build(),
 schemaRegistryConfig);

 RetrievalConfig retrievalConfig =
 configsBuilder.retrievalConfig().retrievalSpecificConfig(new
 PollingConfig(streamName, kinesisClient));
 retrievalConfig.glueSchemaRegistryDeserializer(glueSchemaRegistryDeserializer);

  Scheduler scheduler = new Scheduler(
            configsBuilder.checkpointConfig(),
            configsBuilder.coordinatorConfig(),
            configsBuilder.leaseManagementConfig(),
            configsBuilder.lifecycleConfig(),
            configsBuilder.metricsConfig(),
            configsBuilder.processorConfig(),
            retrievalConfig
```

```
       );

public void processRecords(ProcessRecordsInput processRecordsInput) {
           MDC.put(SHARD_ID_MDC_KEY, shardId);
           try {
               log.info("Processing {} record(s)",
               processRecordsInput.records().size());
               processRecordsInput.records()
               .forEach(
                   r ->
                       log.info("Processed record pk: {} -- Seq: {} : data {} with
schema: {}",
                       r.partitionKey(),
r.sequenceNumber(), recordToAvroObj(r).toString(), r.getSchema()));
           } catch (Throwable t) {
               log.error("Caught throwable while processing records. Aborting.");
               Runtime.getRuntime().halt(1);
           } finally {
               MDC.remove(SHARD_ID_MDC_KEY);
           }
}

private GenericRecord recordToAvroObj(KinesisClientRecord r) {
   byte[] data = new byte[r.data().remaining()];
   r.data().get(data, 0, data.length);
   org.apache.avro.Schema schema = new
org.apache.avro.Schema.Parser().parse(r.schema().getSchemaDefinition());
   DatumReader datumReader = new GenericDatumReader<>(schema);

   BinaryDecoder binaryDecoder = DecoderFactory.get().binaryDecoder(data, 0,
data.length, null);
   return (GenericRecord) datumReader.read(null, binaryDecoder);
}
```

Interacting with data using the Kinesis Data Streams APIs

This section describes integrating Kinesis Data Streams with Schema Registry using the Kinesis Data Streams APIs.

1. Update these Maven dependencies:

```
<dependencyManagement>
```

```
        <dependencies>
            <dependency>
                <groupId>com.amazonaws</groupId>
                <artifactId>aws-java-sdk-bom</artifactId>
                <version>1.11.884</version>
                <type>pom</type>
                <scope>import</scope>
            </dependency>
        </dependencies>
    </dependencyManagement>

    <dependencies>
        <dependency>
            <groupId>com.amazonaws</groupId>
            <artifactId>aws-java-sdk-kinesis</artifactId>
        </dependency>

        <dependency>
            <groupId>software.amazon.glue</groupId>
            <artifactId>schema-registry-serde</artifactId>
            <version>1.1.5</version>
        </dependency>

        <dependency>
            <groupId>com.fasterxml.jackson.dataformat</groupId>
            <artifactId>jackson-dataformat-cbor</artifactId>
            <version>2.11.3</version>
        </dependency>
    </dependencies>
```

2. In the producer, add schema header information using the `PutRecords` or `PutRecord` API in Kinesis Data Streams.

```
//The following lines add a Schema Header to the record
        com.amazonaws.services.schemaregistry.common.Schema awsSchema =
            new com.amazonaws.services.schemaregistry.common.Schema(schemaDefinition,
 DataFormat.AVRO.name(),
                schemaName);
        GlueSchemaRegistrySerializerImpl glueSchemaRegistrySerializer =
            new
 GlueSchemaRegistrySerializerImpl(DefaultCredentialsProvider.builder().build(), new
 GlueSchemaRegistryConfiguration(getConfigs()));
        byte[] recordWithSchemaHeader =
            glueSchemaRegistrySerializer.encode(streamName, awsSchema,
 recordAsBytes);
```

3. In the producer, use the `PutRecords` or `PutRecord` API to put the record into the data stream.
4. In the consumer, remove the schema record from the header, and serialize an Avro schema record.

```
//The following lines remove Schema Header from record
        GlueSchemaRegistryDeserializerImpl glueSchemaRegistryDeserializer =
            new
 GlueSchemaRegistryDeserializerImpl(DefaultCredentialsProvider.builder().build(),
 getConfigs());
        byte[] recordWithSchemaHeaderBytes = new
 byte[recordWithSchemaHeader.remaining()];
        recordWithSchemaHeader.get(recordWithSchemaHeaderBytes, 0,
 recordWithSchemaHeaderBytes.length);
        com.amazonaws.services.schemaregistry.common.Schema awsSchema =
            glueSchemaRegistryDeserializer.getSchema(recordWithSchemaHeaderBytes);
        byte[] record =
 glueSchemaRegistryDeserializer.getData(recordWithSchemaHeaderBytes);

        //The following lines serialize an AVRO schema record
        if (DataFormat.AVRO.name().equals(awsSchema.getDataFormat())) {
            Schema avroSchema = new
 org.apache.avro.Schema.Parser().parse(awsSchema.getSchemaDefinition());
            Object genericRecord = convertBytesToRecord(avroSchema, record);
            System.out.println(genericRecord);
        }
```

Interacting with data using the Kinesis Data Streams APIs

The following is example code for using the `PutRecords` and `GetRecords` APIs.

```
//Full sample code
import
 com.amazonaws.services.schemaregistry.deserializers.GlueSchemaRegistryDeserializerImpl;
import
 com.amazonaws.services.schemaregistry.serializers.GlueSchemaRegistrySerializerImpl;
import com.amazonaws.services.schemaregistry.utils.AVROUtils;
import com.amazonaws.services.schemaregistry.utils.AWSSchemaRegistryConstants;
import org.apache.avro.Schema;
import org.apache.avro.generic.GenericData;
import org.apache.avro.generic.GenericDatumReader;
import org.apache.avro.generic.GenericDatumWriter;
import org.apache.avro.generic.GenericRecord;
import org.apache.avro.io.Decoder;
import org.apache.avro.io.DecoderFactory;
import org.apache.avro.io.Encoder;
import org.apache.avro.io.EncoderFactory;
import software.amazon.awssdk.auth.credentials.DefaultCredentialsProvider;
import software.amazon.awssdk.services.glue.model.DataFormat;

import java.io.ByteArrayOutputStream;
import java.io.File;
import java.io.IOException;
import java.nio.ByteBuffer;
import java.util.Collections;
import java.util.HashMap;
import java.util.Map;


public class PutAndGetExampleWithEncodedData {
    static final String regionName = "us-east-2";
    static final String streamName = "testStream1";
    static final String schemaName = "User-Topic";
    static final String AVRO_USER_SCHEMA_FILE = "src/main/resources/user.avsc";
    KinesisApi kinesisApi = new KinesisApi();

    void runSampleForPutRecord() throws IOException {
        Object testRecord = getTestRecord();
        byte[] recordAsBytes = convertRecordToBytes(testRecord);
        String schemaDefinition =
 AVROUtils.getInstance().getSchemaDefinition(testRecord);

        //The following lines add a Schema Header to a record
        com.amazonaws.services.schemaregistry.common.Schema awsSchema =
            new com.amazonaws.services.schemaregistry.common.Schema(schemaDefinition,
 DataFormat.AVRO.name(),
                schemaName);
        GlueSchemaRegistrySerializerImpl glueSchemaRegistrySerializer =
            new
 GlueSchemaRegistrySerializerImpl(DefaultCredentialsProvider.builder().build(), new
 GlueSchemaRegistryConfiguration(regionName));
        byte[] recordWithSchemaHeader =
            glueSchemaRegistrySerializer.encode(streamName, awsSchema, recordAsBytes);
        //Use PutRecords api to pass a list of records
        kinesisApi.putRecords(Collections.singletonList(recordWithSchemaHeader),
 streamName, regionName);

        //OR
        //Use PutRecord api to pass single record
        //kinesisApi.putRecord(recordWithSchemaHeader, streamName, regionName);
    }

    byte[] runSampleForGetRecord() throws IOException {
        ByteBuffer recordWithSchemaHeader = kinesisApi.getRecords(streamName,
 regionName);

        //The following lines remove the schema registry header
        GlueSchemaRegistryDeserializerImpl glueSchemaRegistryDeserializer =
            new
 GlueSchemaRegistryDeserializerImpl(DefaultCredentialsProvider.builder().build(), new
 GlueSchemaRegistryConfiguration(regionName));
        byte[] recordWithSchemaHeaderBytes = new
 byte[recordWithSchemaHeader.remaining()];
        recordWithSchemaHeader.get(recordWithSchemaHeaderBytes, 0,
 recordWithSchemaHeaderBytes.length);

        com.amazonaws.services.schemaregistry.common.Schema awsSchema =
            glueSchemaRegistryDeserializer.getSchema(recordWithSchemaHeaderBytes);

        byte[] record =
 glueSchemaRegistryDeserializer.getData(recordWithSchemaHeaderBytes);

        //The following lines serialize an AVRO schema record
        if (DataFormat.AVRO.name().equals(awsSchema.getDataFormat())) {
            Schema avroSchema = new
 org.apache.avro.Schema.Parser().parse(awsSchema.getSchemaDefinition());
            Object genericRecord = convertBytesToRecord(avroSchema, record);
            System.out.println(genericRecord);
        }

        return record;
    }

    private byte[] convertRecordToBytes(final Object record) throws IOException {
        ByteArrayOutputStream recordAsBytes = new ByteArrayOutputStream();
        Encoder encoder = EncoderFactory.get().directBinaryEncoder(recordAsBytes,
 null);
```

```
        GenericDatumWriter datumWriter = new
GenericDatumWriter<>(AVROUtils.getInstance().getSchema(record));
        datumWriter.write(record, encoder);
        encoder.flush();
        return recordAsBytes.toByteArray();
    }

    private GenericRecord convertBytesToRecord(Schema avroSchema, byte[] record) throws
IOException {
        final GenericDatumReader<GenericRecord> datumReader = new
GenericDatumReader<>(avroSchema);
        Decoder decoder = DecoderFactory.get().binaryDecoder(record, null);
        GenericRecord genericRecord = datumReader.read(null, decoder);
        return genericRecord;
    }

    private Map<String, String> getMetadata() {
        Map<String, String> metadata = new HashMap<>();
        metadata.put("event-source-1", "topic1");
        metadata.put("event-source-2", "topic2");
        metadata.put("event-source-3", "topic3");
        metadata.put("event-source-4", "topic4");
        metadata.put("event-source-5", "topic5");
        return metadata;
    }

    private GlueSchemaRegistryConfiguration getConfigs() {
        GlueSchemaRegistryConfiguration configs = new
GlueSchemaRegistryConfiguration(regionName);
        configs.setSchemaName(schemaName);
        configs.setAutoRegistration(true);
        configs.setMetadata(getMetadata());
        return configs;
    }

    private Object getTestRecord() throws IOException {
        GenericRecord genericRecord;
        Schema.Parser parser = new Schema.Parser();
        Schema avroSchema = parser.parse(new File(AVRO_USER_SCHEMA_FILE));

        genericRecord = new GenericData.Record(avroSchema);
        genericRecord.put("name", "testName");
        genericRecord.put("favorite_number", 99);
        genericRecord.put("favorite_color", "red");
        return genericRecord;
    }
}
```

Use case: Amazon Managed Service for Apache Flink

Apache Flink is a popular open source framework and distributed processing engine for stateful computations over unbounded and bounded data streams. Amazon Managed Service for Apache Flink is a fully managed AWS service that enables you to build and manage Apache Flink applications to process streaming data.

Open source Apache Flink provides a number of sources and sinks. For example, predefined data sources include reading from files, directories, and sockets, and ingesting data from collections and iterators. Apache Flink DataStream Connectors provide code for Apache Flink to interface with various third-party systems, such as Apache Kafka or Kinesis as sources and/or sinks.

For more information, see Amazon Kinesis Data Analytics Developer Guide.

Apache Flink Kafka connector

Apache Flink provides an Apache Kafka data stream connector for reading data from and writing data to Kafka topics with exactly-once guarantees. Flink's Kafka consumer, `FlinkKafkaConsumer`, provides access to read from one or more Kafka topics. Apache Flink’s Kafka Producer, `FlinkKafkaProducer`, allows writing a stream of records to one or more Kafka topics. For more information, see Apache Kafka Connector.

Apache Flink Kinesis streams Connector

The Kinesis data stream connector provides access to Amazon Kinesis Data Streams. The `FlinkKinesisConsumer` is an exactly-once parallel streaming data source that subscribes to multiple Kinesis streams within the same AWS service region, and can transparently handle re-sharding of streams while the job is running. Each subtask of the consumer is responsible for fetching data records from multiple Kinesis shards. The number of shards fetched by each subtask will change as shards are closed and created by Kinesis. The `FlinkKinesisProducer` uses Kinesis Producer Library (KPL) to put data from an Apache Flink stream into a Kinesis stream. For more information, see Amazon Kinesis Streams Connector.

For more information, see the AWS Glue Schema Github repository.

Integrating with Apache Flink

The SerDes library provided with Schema Registry integrates with Apache Flink. To work with Apache Flink, you are required to implement `SerializationSchema` and `DeserializationSchema` interfaces called `GlueSchemaRegistryAvroSerializationSchema` and `GlueSchemaRegistryAvroDeserializationSchema`, which you can plug into Apache Flink connectors.

Adding an AWS Glue Schema Registry dependency into the Apache Flink application

To set up the integration dependencies to AWS Glue Schema Registry in the Apache Flink application:

1. Add the dependency to the `pom.xml` file.

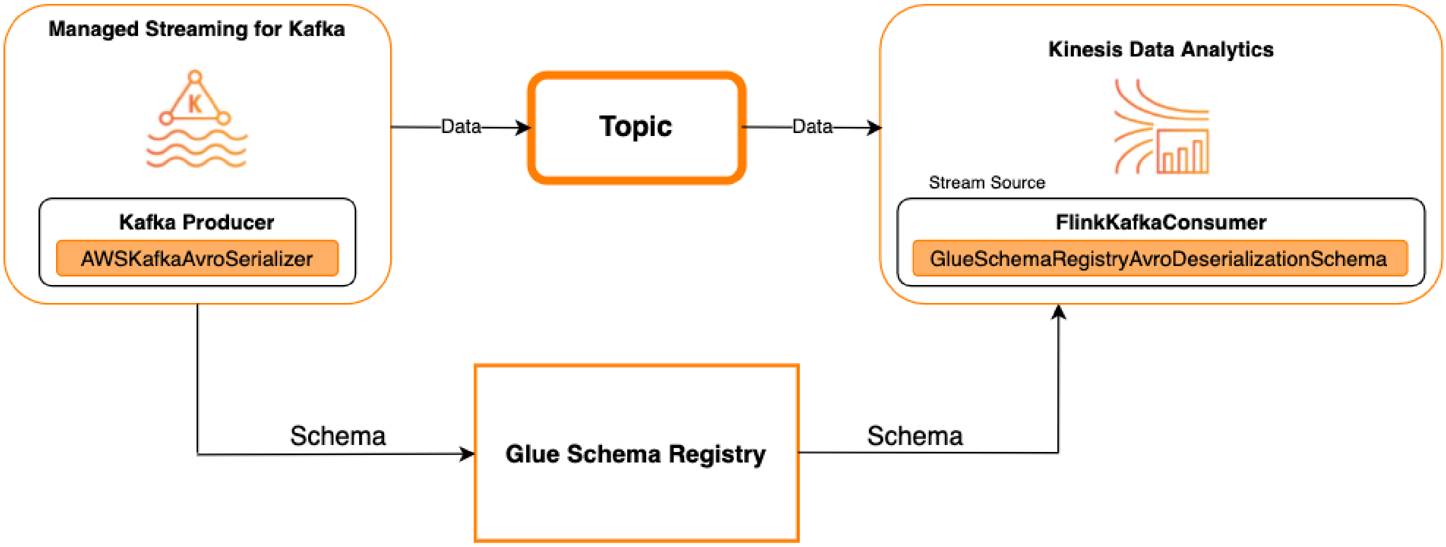
```
<dependency>
    <groupId>software.amazon.glue</groupId>
    <artifactId>schema-registry-flink-serde</artifactId>
    <version>1.0.0</version>
</dependency>
```

Integrating Kafka or Amazon MSK with Apache Flink

You can use Managed Service for Apache Flink for Apache Flink, with Kafka as a source or Kafka as a sink.

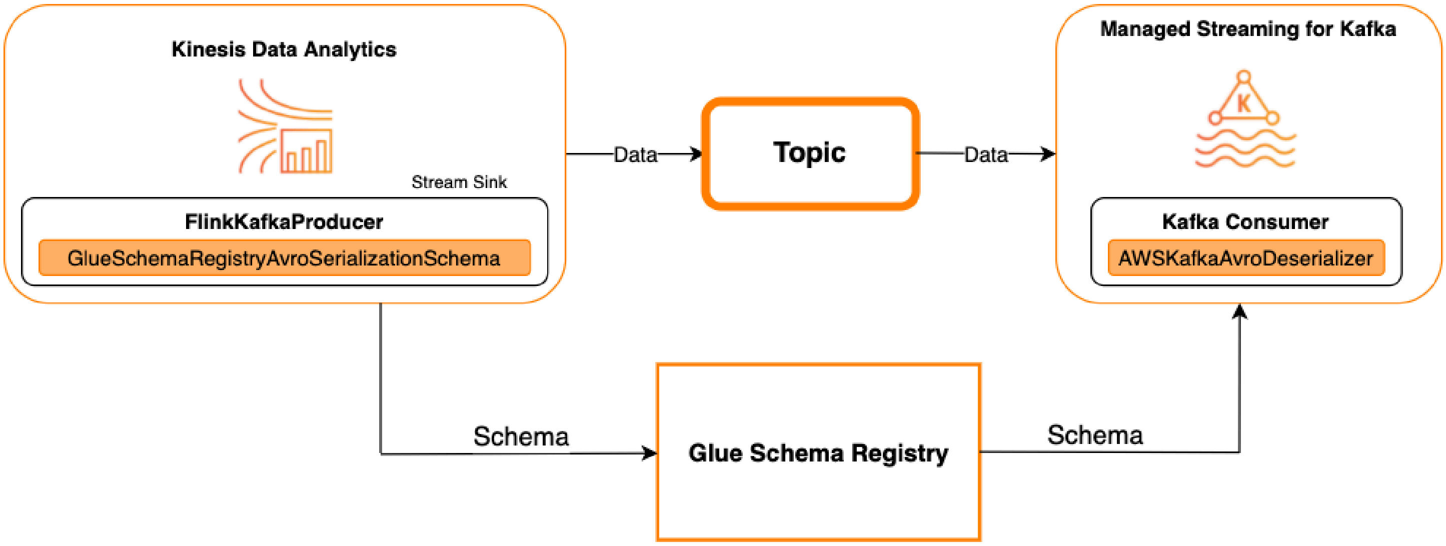
Kafka as a source

The following diagram shows integrating Kinesis Data Streams with Managed Service for Apache Flink for Apache Flink, with Kafka as a source.


Managed Streaming for Kafka
Kafka Producer
AWSKafkaAvroSerializer
Data
Topic
Data
Kinesis Data Analytics
Stream Source
FlinkKafkaConsumer
GlueSchemaRegistryAvroDeserializationSchema
Schema
Glue Schema Registry
Schema

Kafka as a sink

The following diagram shows integrating Kinesis Data Streams with Managed Service for Apache Flink for Apache Flink, with Kafka as a sink.


Kinesis Data Analytics
Stream Sink
FlinkKafkaProducer
GlueSchemaRegistryAvroSerializationSchema
Data
Topic
Data
Managed Streaming for Kafka
Kafka Consumer
AWSKafkaAvroDeserializer
Schema
Glue Schema Registry
Schema

To integrate Kafka (or Amazon MSK) with Managed Service for Apache Flink for Apache Flink, with Kafka as a source or Kafka as a sink, make the code changes below. Add the bolded code blocks to your respective code in the analogous sections.

If Kafka is the source, then use the deserializer code (block 2). If Kafka is the sink, use the serializer code (block 3).

```
StreamExecutionEnvironment env = StreamExecutionEnvironment.getExecutionEnvironment();
```

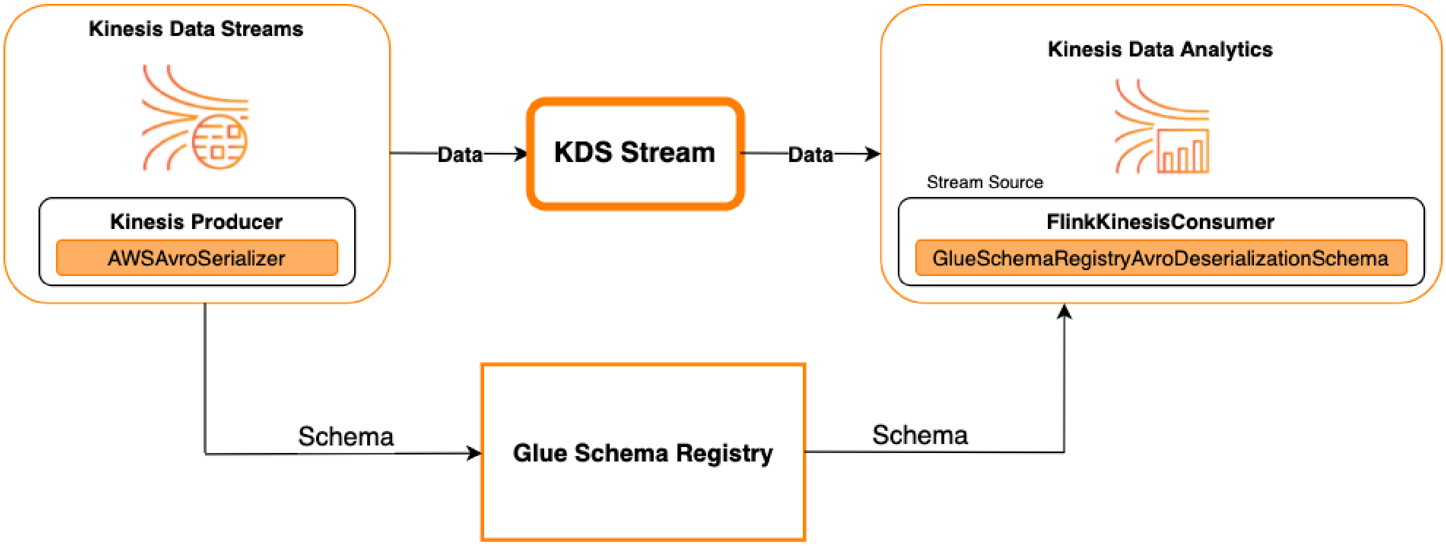
```
String topic = "topic";
Properties properties = new Properties();
properties.setProperty("bootstrap.servers", "localhost:9092");
properties.setProperty("group.id", "test");

// block 1
Map<String, Object> configs = new HashMap<>();
configs.put(AWSSchemaRegistryConstants.AWS_REGION, "aws-region");
configs.put(AWSSchemaRegistryConstants.SCHEMA_AUTO_REGISTRATION_SETTING, true);
configs.put(AWSSchemaRegistryConstants.AVRO_RECORD_TYPE,
 AvroRecordType.GENERIC_RECORD.getName());

FlinkKafkaConsumer<GenericRecord> consumer = new FlinkKafkaConsumer<>(
    topic,
    // block 2
    GlueSchemaRegistryAvroDeserializationSchema.forGeneric(schema, configs),
    properties);

FlinkKafkaProducer<GenericRecord> producer = new FlinkKafkaProducer<>(
    topic,
    // block 3
    GlueSchemaRegistryAvroSerializationSchema.forGeneric(schema, topic, configs),
    properties);

DataStream<GenericRecord> stream = env.addSource(consumer);
stream.addSink(producer);
env.execute();
```

Integrating Kinesis Data Streams with Apache Flink

You can use Managed Service for Apache Flink for Apache Flink with Kinesis Data Streams as a source or a sink.

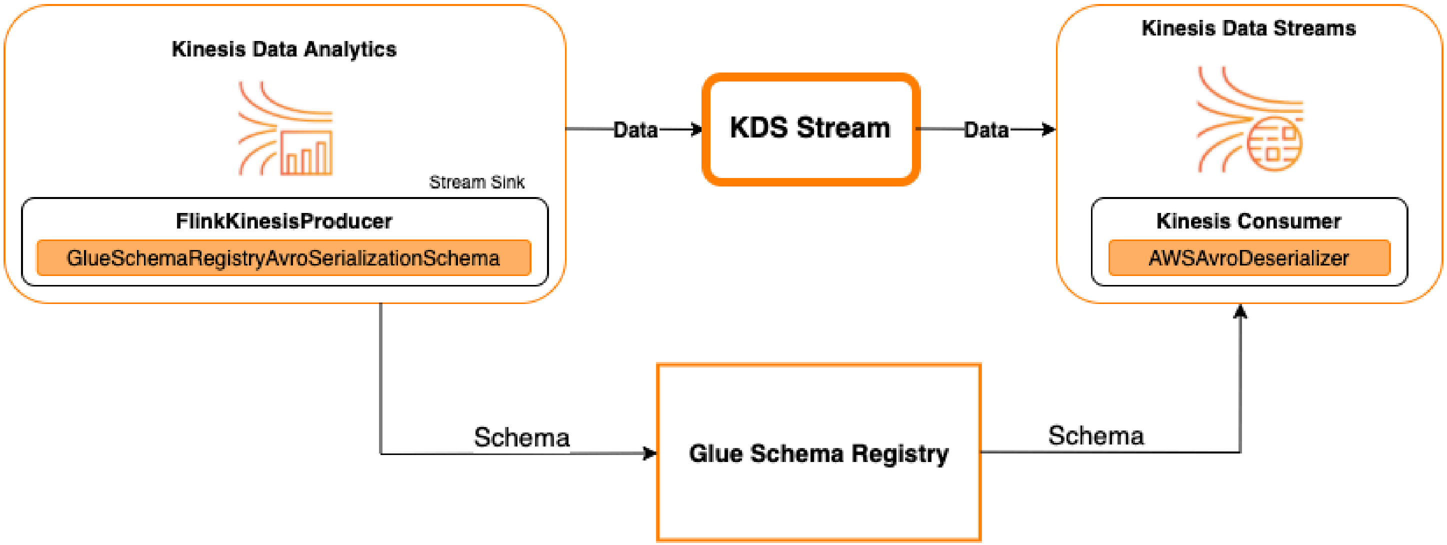
Kinesis Data Streams as a source

The following diagram shows integrating Kinesis Data Streams with Managed Service for Apache Flink for Apache Flink, with Kinesis Data Streams as a source.


Kinesis Data Streams
Kinesis Producer
AWSAvroSerializer
Data
KDS Stream
Data
Kinesis Data Analytics
Stream Source
FlinkKinesisConsumer
GlueSchemaRegistryAvroDeserializationSchema
Schema
Glue Schema Registry
Schema

Kinesis Data Streams as a sink

The following diagram shows integrating Kinesis Data Streams with Managed Service for Apache Flink for Apache Flink, with Kinesis Data Streams as a sink.


Kinesis Data Analytics
Stream Sink
FlinkKinesisProducer
GlueSchemaRegistryAvroSerializationSchema
Data
KDS Stream
Data
Kinesis Data Streams
Kinesis Consumer
AWSAvroDeserializer
Schema
Glue Schema Registry
Schema

To integrate Kinesis Data Streams with Managed Service for Apache Flink for Apache Flink, with Kinesis Data Streams as a source or Kinesis Data Streams as a sink, make the code changes below. Add the bolded code blocks to your respective code in the analogous sections.

If Kinesis Data Streams is the source, use the deserializer code (block 2). If Kinesis Data Streams is the sink, use the serializer code (block 3).

```
StreamExecutionEnvironment env = StreamExecutionEnvironment.getExecutionEnvironment();
```

```
String streamName = "stream";
Properties consumerConfig = new Properties();
consumerConfig.put(AWSConfigConstants.AWS_REGION, "aws-region");
consumerConfig.put(AWSConfigConstants.AWS_ACCESS_KEY_ID, "aws_access_key_id");
consumerConfig.put(AWSConfigConstants.AWS_SECRET_ACCESS_KEY, "aws_secret_access_key");
consumerConfig.put(ConsumerConfigConstants.STREAM_INITIAL_POSITION, "LATEST");

// block 1
Map<String, Object> configs = new HashMap<>();
configs.put(AWSSchemaRegistryConstants.AWS_REGION, "aws-region");
configs.put(AWSSchemaRegistryConstants.SCHEMA_AUTO_REGISTRATION_SETTING, true);
configs.put(AWSSchemaRegistryConstants.AVRO_RECORD_TYPE,
 AvroRecordType.GENERIC_RECORD.getName());

FlinkKinesisConsumer<GenericRecord> consumer = new FlinkKinesisConsumer<>(
    streamName,
    // block 2
    GlueSchemaRegistryAvroDeserializationSchema.forGeneric(schema, configs),
    properties);

FlinkKinesisProducer<GenericRecord> producer = new FlinkKinesisProducer<>(
    // block 3
    GlueSchemaRegistryAvroSerializationSchema.forGeneric(schema, topic, configs),
    properties);
producer.setDefaultStream(streamName);
producer.setDefaultPartition("0");

DataStream<GenericRecord> stream = env.addSource(consumer);
stream.addSink(producer);
env.execute();
```

Use Case: Integration with AWS Lambda

To use an AWS Lambdafunction as an Apache Kafka/Amazon MSK consumer and deserialize Avro-encoded messages using AWS Glue Schema Registry, visit the MSK Labs page.

Use case: AWS Glue Data Catalog

AWS Glue tables support schemas that you can specify manually or by reference to the AWS Glue Schema Registry. The Schema Registry integrates with the Data Catalog to allow you to optionally use schemas stored in the Schema Registry when creating or updating AWS Glue tables or partitions in the Data Catalog. To identify a schema definition in the Schema Registry, at a minimum, you need to know the ARN of the schema it is part of. A schema version of a schema, which contains a schema definition, can be referenced by its UUID or version number. There is always one schema version, the "latest" version, that can be looked up without knowing its version number or UUID.

When calling the `CreateTable` or `UpdateTable` operations, you will pass a `TableInput` structure that contains a `StorageDescriptor`, which may have a `SchemaReference` to an existing schema in the Schema Registry. Similarly, when you call the `GetTable` or `GetPartition` APIs, the response may contain the schema and the `SchemaReference`. When a table or partition was created using a schema references, the Data Catalog will try to fetch the schema for this schema reference. In case it is unable to find the schema in the Schema Registry, it returns an empty schema in the `GetTable` response; otherwise the response will have both the schema and schema reference.

You can also perform the actions from the AWS Glue console.

To perform these operations and create, update, or view the schema information, you must give an IAM role to the calling user that provides permissions for the `GetSchemaVersion` API.

Adding a table or updating the schema for a table

Adding a new table from an existing schema binds the table to a specific schema version. Once new schema versions get registered, you can update this table definition from the View table page in the AWS Glue console or using the UpdateTable action (Python: update_table) API.

Adding a table from an existing schema

You can create an AWS Glue table from a schema version in the registry using the AWS Glue console or `CreateTable` API.

AWS Glue API

When calling the `CreateTable` API, you will pass a `TableInput` that contains a `StorageDescriptor` which has a `SchemaReference` to an existing schema in the Schema Registry.

AWS Glue console

To create a table from the AWS Glue console:

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, under **Data catalog**, choose **Tables**.

3. In the **Add Tables** menu, choose **Add table from existing schema**.
4. Configure the table properties and data store per the AWS Glue Developer Guide.
5. In the **Choose a Glue schema** page, select the **Registry** where the schema resides.
6. Choose the **Schema name** and select the **Version** of the schema to apply.
7. Review the schema preview, and choose **Next**.
8. Review and create the table.

The schema and version applied to the table appears in the **Glue schema** column in the list of tables. You can view the table to see more details.

Updating the schema for a table

When a new schema version becomes available, you may want to update a table's schema using the UpdateTable action (Python: update_table) API or the AWS Glue console.

> ⚠ **Important**
>
> When updating the schema for an existing table that has an AWS Glue schema specified manually, the new schema referenced in the Schema Registry may be incompatible. This can result in your jobs failing.

AWS Glue API

When calling the `UpdateTable` API, you will pass a `TableInput` that contains a `StorageDescriptor` which has a `SchemaReference` to an existing schema in the Schema Registry.

AWS Glue console

To update the schema for a table from the AWS Glue console:

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, under **Data catalog**, choose **Tables**.
3. View the table from the list of tables.
4. Click **Update schema** in the box that informs you about a new version.
5. Review the differences between the current and new schema.

6. Choose **Show all schema differences** to see more details.
7. Choose **Save table** to accept the new version.

Use case: AWS Glue streaming

AWS Glue streaming consumes data from streaming sources and perform ETL operations before writing to an output sink. Input streaming source can be specified using a Data Table or directly by specifying the source configuration.

AWS Glue streaming supports a Data Catalog table for the streaming source created with the schema present in the AWS Glue Schema Registry. You can create a schema in the AWS Glue Schema Registry and create an AWS Glue table with a streaming source using this schema. This AWS Glue table can be used as an input to an AWS Glue streaming job for deserializing data in the input stream.

One point to note here is when the schema in the AWS Glue Schema Registry changes, you need to restart the AWS Glue streaming job needs to reflect the changes in the schema.

Use case: Apache Kafka Streams

The Apache Kafka Streams API is a client library for processing and analyzing data stored in Apache Kafka. This section describes the integration of Apache Kafka Streams with AWS Glue Schema Registry, which allows you to manage and enforce schemas on your data streaming applications. For more information on Apache Kafka Streams, see Apache Kafka Streams.

Integrating with the SerDes Libraries

There is a `GlueSchemaRegistryKafkaStreamsSerde` class that you can configure a Streams application with.

Kafka Streams application example code

To use the AWS Glue Schema Registry within an Apache Kafka Streams application:

1. Configure the Kafka Streams application.

```
final Properties props = new Properties();
    props.put(StreamsConfig.APPLICATION_ID_CONFIG, "avro-streams");
    props.put(StreamsConfig.BOOTSTRAP_SERVERS_CONFIG, "localhost:9092");
    props.put(StreamsConfig.CACHE_MAX_BYTES_BUFFERING_CONFIG, 0);
    props.put(StreamsConfig.DEFAULT_KEY_SERDE_CLASS_CONFIG,
 Serdes.String().getClass().getName());
    props.put(StreamsConfig.DEFAULT_VALUE_SERDE_CLASS_CONFIG,
 AWSKafkaAvroSerDe.class.getName());
    props.put(ConsumerConfig.AUTO_OFFSET_RESET_CONFIG, "earliest");

    props.put(AWSSchemaRegistryConstants.AWS_REGION, "aws-region");
    props.put(AWSSchemaRegistryConstants.SCHEMA_AUTO_REGISTRATION_SETTING, true);
    props.put(AWSSchemaRegistryConstants.AVRO_RECORD_TYPE,
 AvroRecordType.GENERIC_RECORD.getName());
 props.put(AWSSchemaRegistryConstants.DATA_FORMAT, DataFormat.AVRO.name());
```

2. Create a stream from the topic avro-input.

```
StreamsBuilder builder = new StreamsBuilder();
final KStream<String, GenericRecord> source = builder.stream("avro-input");
```

3. Process the data records (the example filters out those records whose value of favorite_color is pink or where the value of amount is 15).

```
final KStream<String, GenericRecord> result = source
    .filter((key, value) -
> !"pink".equals(String.valueOf(value.get("favorite_color"))));
    .filter((key, value) -> !"15.0".equals(String.valueOf(value.get("amount"))));
```

4. Write the results back to the topic avro-output.

```
result.to("avro-output");
```

5. Start the Apache Kafka Streams application.

```
KafkaStreams streams = new KafkaStreams(builder.build(), props);
streams.start();
```

Implementation results

These results show the filtering process of records that were filtered out in step 3 as a favorite_color of "pink" or value of "15.0".

Records before filtering:

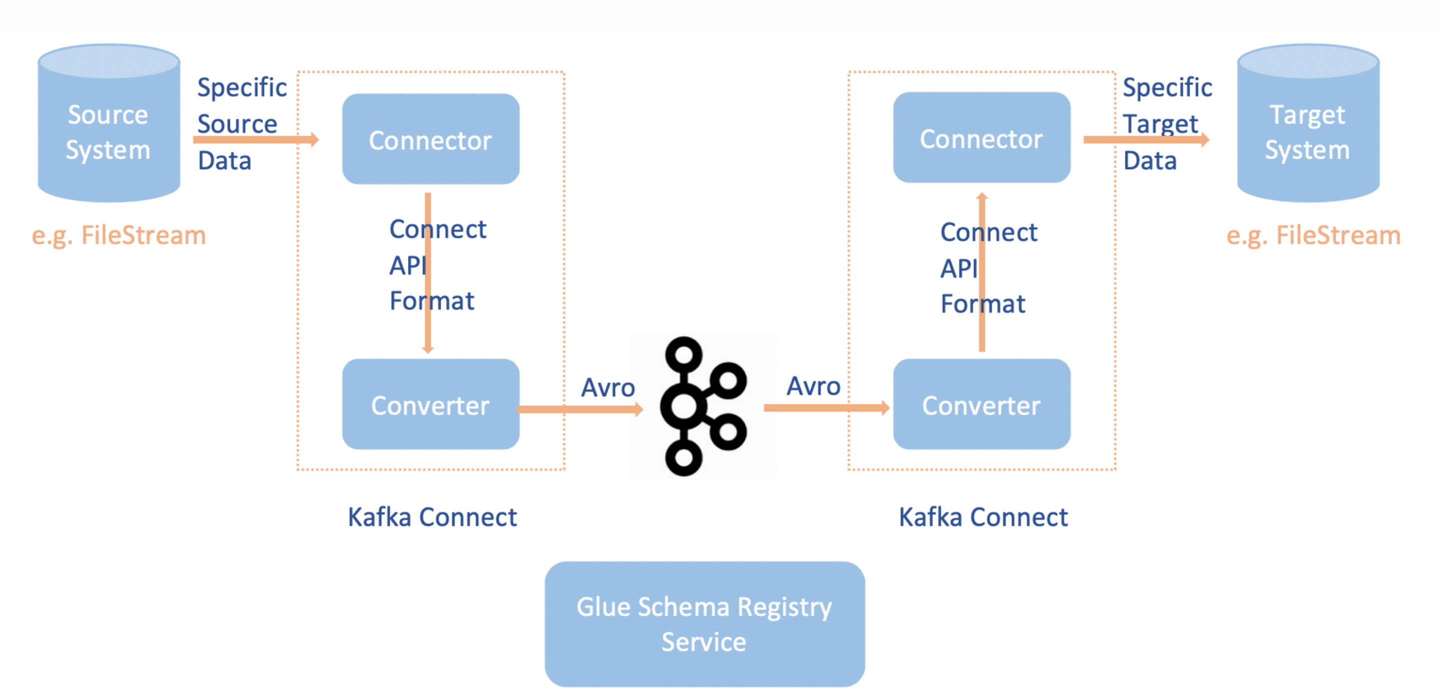
```
{"name": "Sansa", "favorite_number": 99, "favorite_color": "white"}
{"name": "Harry", "favorite_number": 10, "favorite_color": "black"}
{"name": "Hermione", "favorite_number": 1, "favorite_color": "red"}
{"name": "Ron", "favorite_number": 0, "favorite_color": "pink"}
{"name": "Jay", "favorite_number": 0, "favorite_color": "pink"}

{"id": "commute_1","amount": 3.5}
{"id": "grocery_1","amount": 25.5}
{"id": "entertainment_1","amount": 19.2}
{"id": "entertainment_2","amount": 105}
 {"id": "commute_1","amount": 15}
```

Records after filtering:

```
{"name": "Sansa", "favorite_number": 99, "favorite_color": "white"}
{"name": "Harry", "favorite_number": 10, "favorite_color": "black"}
{"name": "Hermione", "favorite_number": 1, "favorite_color": "red"}
{"name": "Ron", "favorite_number": 0, "favorite_color": "pink"}

{"id": "commute_1","amount": 3.5}
{"id": "grocery_1","amount": 25.5}
{"id": "entertainment_1","amount": 19.2}
{"id": "entertainment_2","amount": 105}
```

Use case: Apache Kafka Connect

The integration of Apache Kafka Connect with the AWS Glue Schema Registry enables you to get schema information from connectors. The Apache Kafka converters specify the format of data within Apache Kafka and how to translate it into Apache Kafka Connect data. Every Apache Kafka Connect user will need to configure these converters based on the format they want their data in when loaded from or stored into Apache Kafka. In this way, you can define your own converters to translate Apache Kafka Connect data into the type used in the AWS Glue Schema Registry (for example: Avro) and utilize our serializer to register its schema and do serialization. Then converters are also able to use our deserializer to deserialize data received from Apache Kafka and convert it back into Apache Kafka Connect data. An example workflow diagram is given below.


Source System
Specific Source Data
e.g. FileStream
Connector
Connect API Format
Converter
Avro
Kafka Connect
Avro
Connector
Connect API Format
Converter
Kafka Connect
Specific Target Data
Target System
e.g. FileStream
Glue Schema Registry Service

1. Install the `aws-glue-schema-registry` project by cloning the [Github repository for the AWS Glue Schema Registry](https://github.com/awslabs/aws-glue-schema-registry).

```
git clone git@github.com:awslabs/aws-glue-schema-registry.git
cd aws-glue-schema-registry
mvn clean install
mvn dependency:copy-dependencies
```

2. If you plan on using Apache Kafka Connect in *Standalone* mode, update **connect-standalone.properties** using the instructions below for this step. If you plan on using Apache Kafka Connect in *Distributed* mode, update **connect-avro-distributed.properties** using the same instructions.

    a. Add these properties also to the Apache Kafka connect properties file:

    ```
    key.converter.region=aws-region
    value.converter.region=aws-region
    key.converter.schemaAutoRegistrationEnabled=true
    value.converter.schemaAutoRegistrationEnabled=true
    key.converter.avroRecordType=GENERIC_RECORD
    ```

```
value.converter.avroRecordType=GENERIC_RECORD
```

b. Add the command below to the **Launch mode** section under **kafka-run-class.sh**:

```
-cp $CLASSPATH:"<your AWS GlueSchema Registry base directory>/target/dependency/*"
```

3. Add the command below to the **Launch mode** section under **kafka-run-class.sh**

```
-cp $CLASSPATH:"<your AWS GlueSchema Registry base directory>/target/dependency/*"
```

It should look like this:

```

# Launch mode
if [ "x$DAEMON_MODE" = "xtrue" ]; then
  nohup "$JAVA" $KAFKA_HEAP_OPTS $KAFKA_JVM_PERFORMANCE_OPTS $KAFKA_GC_LOG_OPTS
 $KAFKA_JMX_OPTS $KAFKA_LOG4J_OPTS -cp $CLASSPATH:"/Users/johndoe/aws-glue-schema-
registry/target/dependency/*" $KAFKA_OPTS "$@" > "$CONSOLE_OUTPUT_FILE" 2>&1 < /dev/
null &
else
  exec "$JAVA" $KAFKA_HEAP_OPTS $KAFKA_JVM_PERFORMANCE_OPTS $KAFKA_GC_LOG_OPTS
 $KAFKA_JMX_OPTS $KAFKA_LOG4J_OPTS -cp $CLASSPATH:"/Users/johndoe/aws-glue-schema-
registry/target/dependency/*" $KAFKA_OPTS "$@"
fi
```

4. If using bash, run the below commands to set-up your CLASSPATH in your bash_profile. For any other shell, update the environment accordingly.

```
echo 'export GSR_LIB_BASE_DIR=<>' >>~/.bash_profile
echo 'export GSR_LIB_VERSION=1.0.0' >>~/.bash_profile
echo 'export KAFKA_HOME=<your Apache Kafka installation directory>' >>~/.bash_profile
echo 'export CLASSPATH=$CLASSPATH:$GSR_LIB_BASE_DIR/avro-kafkaconnect-converter/
target/schema-registry-kafkaconnect-converter-$GSR_LIB_VERSION.jar:$GSR_LIB_BASE_DIR/
common/target/schema-registry-common-$GSR_LIB_VERSION.jar:$GSR_LIB_BASE_DIR/
avro-serializer-deserializer/target/schema-registry-serde-$GSR_LIB_VERSION.jar'
 >>~/.bash_profile
source ~/.bash_profile
```

5. (Optional) If you want to test with a simple file source, then clone the file source connector.

```
git clone https://github.com/mmolimar/kafka-connect-fs.git
cd kafka-connect-fs/
```

   a. Under the source connector configuration, edit the data format to Avro, file reader to `AvroFileReader` and update an example Avro object from the file path you are reading from. For example:

      ```
      vim config/kafka-connect-fs.properties
      ```

      ```
      fs.uris=<path to a sample avro object>
      policy.regexp=^.*\.avro$
      file_reader.class=com.github.mmolimar.kafka.connect.fs.file.reader.AvroFileReader
      ```

   b. Install the source connector.

      ```
      mvn clean package
      echo "export CLASSPATH=\$CLASSPATH:\"\$(find target/ -type f -name '*.jar'| grep
       '\-package' | tr '\n' ':')\"" >>~/.bash_profile
      source ~/.bash_profile
      ```

   c. Update the sink properties under *<your Apache Kafka installation directory>*/`config/connect-file-sink.properties` update the topic name and out file name.

      ```
      file=<output file full path>
      topics=<my topic>
      ```

6. Start the Source Connector (in this example it is a file source connector).

   ```
   $KAFKA_HOME/bin/connect-standalone.sh $KAFKA_HOME/config/connect-
   standalone.properties config/kafka-connect-fs.properties
   ```

7. Run the Sink Connector (in this example it is a file sink connector).

   ```
   $KAFKA_HOME/bin/connect-standalone.sh $KAFKA_HOME/config/connect-
   standalone.properties $KAFKA_HOME/config/connect-file-sink.properties
   ```

   For an example Kafka Connect usage, look at the run-local-tests.sh script under integration-tests folder in the Github repository for the AWS Glue Schema Registry.

Migration from a third-party schema registry to AWS Glue Schema Registry

The migration from a third-party schema registry to the AWS Glue Schema Registry has a dependency on the existing, current third-party schema registry. If there are records in an Apache Kafka topic which were sent using a third-party schema registry, consumers need the third-party schema registry to deserialize those records. The `AWSKafkaAvroDeserializer` provides the ability to specify a secondary deserializer class which points to the third-party deserializer and is used to deserialize those records.

There are two criteria for retirement of a third-party schema. First, retirement can occur only after records in Apache Kafka topics using the 3rd party schema registry are either no longer required by and for any consumers. Second, retirement can occur by aging out of the Apache Kafka topics, depending on the retention period specified for those topics. Note that if you have topics which have infinite retention, you can still migrate to the AWS Glue Schema Registry but you will not be able to retire the third-party schema registry. As a workaround, you can use an application or Mirror Maker 2 to read from the current topic and produce to a new topic with the AWS Glue Schema Registry.

To migrate from a third-party schema registry to the AWS Glue Schema Registry:

1. Create a registry in the AWS Glue Schema Registry, or use the default registry.
2. Stop the consumer. Modify it to include AWS Glue Schema Registry as the primary deserializer, and the third-party schema registry as the secondary.
    - Set the consumer properties. In this example, the secondary_deserializer is set to a different deserializer. The behavior is as follows: the consumer retrieves records from Amazon MSK and first tries to use the `AWSKafkaAvroDeserializer`. If it is unable to read the magic byte that contains the Avro Schema ID for the AWS Glue Schema Registry schema, the `AWSKafkaAvroDeserializer` then tries to use the deserializer class provided in the secondary_deserializer. The properties specific to the secondary deserializer also need to be provided in the consumer properties, such as the schema_registry_url_config and specific_avro_reader_config, as shown below.

```
consumerProps.setProperty(ConsumerConfig.KEY_DESERIALIZER_CLASS_CONFIG,
 StringDeserializer.class.getName());
consumerProps.setProperty(ConsumerConfig.VALUE_DESERIALIZER_CLASS_CONFIG,
 AWSKafkaAvroDeserializer.class.getName());
```

```
consumerProps.setProperty(AWSSchemaRegistryConstants.AWS_REGION,
 KafkaClickstreamConsumer.gsrRegion);
consumerProps.setProperty(AWSSchemaRegistryConstants.SECONDARY_DESERIALIZER,
 KafkaAvroDeserializer.class.getName());
consumerProps.setProperty(KafkaAvroDeserializerConfig.SCHEMA_REGISTRY_URL_CONFIG,
 "URL for third-party schema registry");
consumerProps.setProperty(KafkaAvroDeserializerConfig.SPECIFIC_AVRO_READER_CONFIG,
 "true");
```

3. Restart the consumer.
4. Stop the producer and point the producer to the AWS Glue Schema Registry.
    a. Set the producer properties. In this example, the producer will use the default-registry and auto register schema versions.

    ```
    producerProps.setProperty(ProducerConfig.KEY_SERIALIZER_CLASS_CONFIG,
     StringSerializer.class.getName());
    producerProps.setProperty(ProducerConfig.VALUE_SERIALIZER_CLASS_CONFIG,
     AWSKafkaAvroSerializer.class.getName());
    producerProps.setProperty(AWSSchemaRegistryConstants.AWS_REGION, "us-east-2");
    producerProps.setProperty(AWSSchemaRegistryConstants.AVRO_RECORD_TYPE,
     AvroRecordType.SPECIFIC_RECORD.getName());
    producerProps.setProperty(AWSSchemaRegistryConstants.SCHEMA_AUTO_REGISTRATION_SETTING,
     "true");
    ```

5. (Optional) Manually move existing schemas and schema versions from the current third-party schema registry to the AWS Glue Schema Registry, either to the default-registry in AWS Glue Schema Registry or to a specific non-default registry in AWS Glue Schema Registry. This can be done by exporting schemas from the third-party schema registries in JSON format and creating new schemas in AWS Glue Schema Registry using the AWS Management Console or the AWS CLI.

    This step may be important if you need to enable compatibility checks with previous schema versions for newly created schema versions using the AWS CLI and the AWS Management Console, or when producers send messages with a new schema with auto-registration of schema versions turned on.
6. Start the producer.

Connecting to data

An AWS Glue *connection* is a Data Catalog object that stores login credentials, URI strings, virtual private cloud (VPC) information, and more for a particular data store. AWS Glue crawlers, jobs, and development endpoints use connections in order to access certain types of data stores. You can use connections for both sources and targets, and reuse the same connection across multiple crawler or extract, transform, and load (ETL) jobs.

The latest version of the AWS Glue connections schema provides a unified way to manage data connections across AWS services and applications, such as AWS Glue, Amazon Athena, and Amazon SageMaker AI Unified Studio.

Overview of using connectors and connections

A *connection* contains the properties that are required to connect to a particular data store. When you create a connection, it is stored in the AWS Glue Data Catalog. You choose a connector, and then create a connection based on that connector.

You can subscribe to connectors for non-natively supported data stores in AWS Marketplace, and then use those connectors when you're creating connections. Developers can also create their own connectors, and you can use them when creating connections.

> **Note**
>
> Connections created using custom or AWS Marketplace connectors in AWS Glue Studio appear in the AWS Glue console with type set to `UNKNOWN`.

The following steps describe the overall process of using connectors in AWS Glue Studio:

1. Subscribe to a connector in AWS Marketplace, or develop your own connector and upload it to AWS Glue Studio. For more information, see Adding connectors to AWS Glue Studio.
2. Review the connector usage information. You can find this information on the **Usage** tab on the connector product page. For example, if you click the **Usage** tab on this product page, AWS Glue Connector for Google BigQuery, you can see in the **Additional Resources** section a link to a blog about using this connector.

3. Create a connection. You choose which connector to use and provide additional information for the connection, such as login credentials, URI strings, and virtual private cloud (VPC) information. For more information, see Creating connections for connectors.
4. Create an IAM role for your job. The job assumes the permissions of the IAM role that you specify when you create it. This IAM role must have the necessary permissions to authenticate with, extract data from, and write data to your data stores.
5. Create an ETL job and configure the data source properties for your ETL job. Provide the connection options and authentication information as instructed by the custom connector provider. For more information, see Authoring jobs with custom connectors.
6. Customize your ETL job by adding transforms or additional data stores, as described in Starting visual ETL jobs in AWS Glue Studio.
7. If using a connector for the data target, configure the data target properties for your ETL job. Provide the connection options and authentication information as instructed by the custom connector provider. For more information, see the section called “Authoring jobs with custom connectors”.
8. Customize the job run environment by configuring job properties, as described in Modify the job properties.
9. Run the job.

Unified connections

AWS recently introduced a new feature called "SageMaker LakeHouse Connections" or "AWS Glue Unified Connections." This feature allows you to create connections that can be used by multiple AWS services, such as AWS Glue and Amazon Athena. When you create a data source in Amazon Athena, you'll notice a section that refers to AWS Glue connection inputs. In this case, Amazon Athena will create a AWS Glue connection for you, including any Amazon Athena-specific properties in the `AthenaProperties` section of the connection.

On the other hand, if you create a connection directly in AWS Glue, you'll only be prompted to enter properties specific to AWS Glue and Apache Spark, which will be stored in the `ConnectionProperties` and `SparkProperties` sections of the connection.

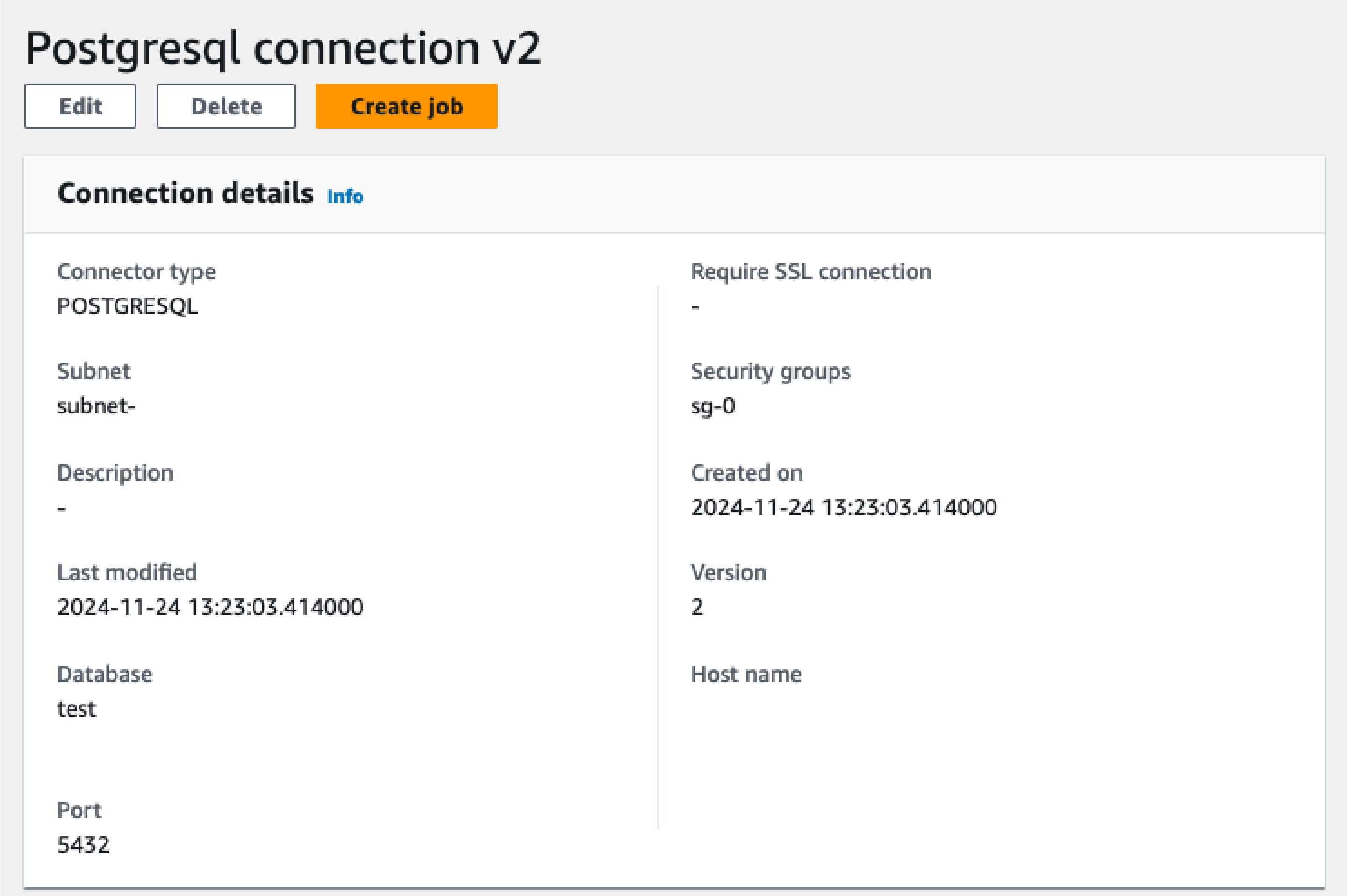
Both of these scenarios result in the creation of a "unified connection," but the connections created in Amazon Athena are only configured for use within Amazon Athena, while the connections created in AWS Glue are only configured for use within AWS Glue. However, it's possible to update these connections with the missing properties (either Amazon Athena or Spark properties) so that they can be used by both services. Amazon SageMaker AI Unified Studio takes care of this automatically by filling in all necessary properties (`` `ConnectionProperties` ``, `` `AthenaProperties` ``, and `` `SparkProperties` ``) on the AWS Glue connection, ensuring that the connection can be used by both AWS Glue and Amazon Athena.

It's important to note that although we refer to these as "unified connections," the connections created in AWS Glue or Amazon Athena individually are not truly unified unless they are properly configured for use by both services. Only the connections created through SageMaker Unified Studio are truly unified and usable by multiple services out of the box.

Additionally, connections created in AWS Glue are not visible in Amazon Athena because Amazon Athena displays data sources, which include a reference to a AWS Glue connection but are not the AWS Glue connection itself. Similarly, connections created in Amazon Athena are not visible in AWS Glue Studio because AWS Glue Studio filters out any connection that hasn't been configured with the necessary settings for AWS Glue.

AWS Glue Studio creates unified connections by default. In the AWS Glue console, you can see the version of the connection in the connections table on the connections page, on the connections detail page, and the connections table in the job details page.

The connection version is visible on Connection details:


Postgresql connection v2
Edit
Delete
Create job
Connection details
Info

Connector type
POSTGRESQL

Require SSL connection
-

Subnet
subnet-

Security groups
sg-0

Description
-

Created on
2024-11-24 13:23:03.414000

Last modified
2024-11-24 13:23:03.414000

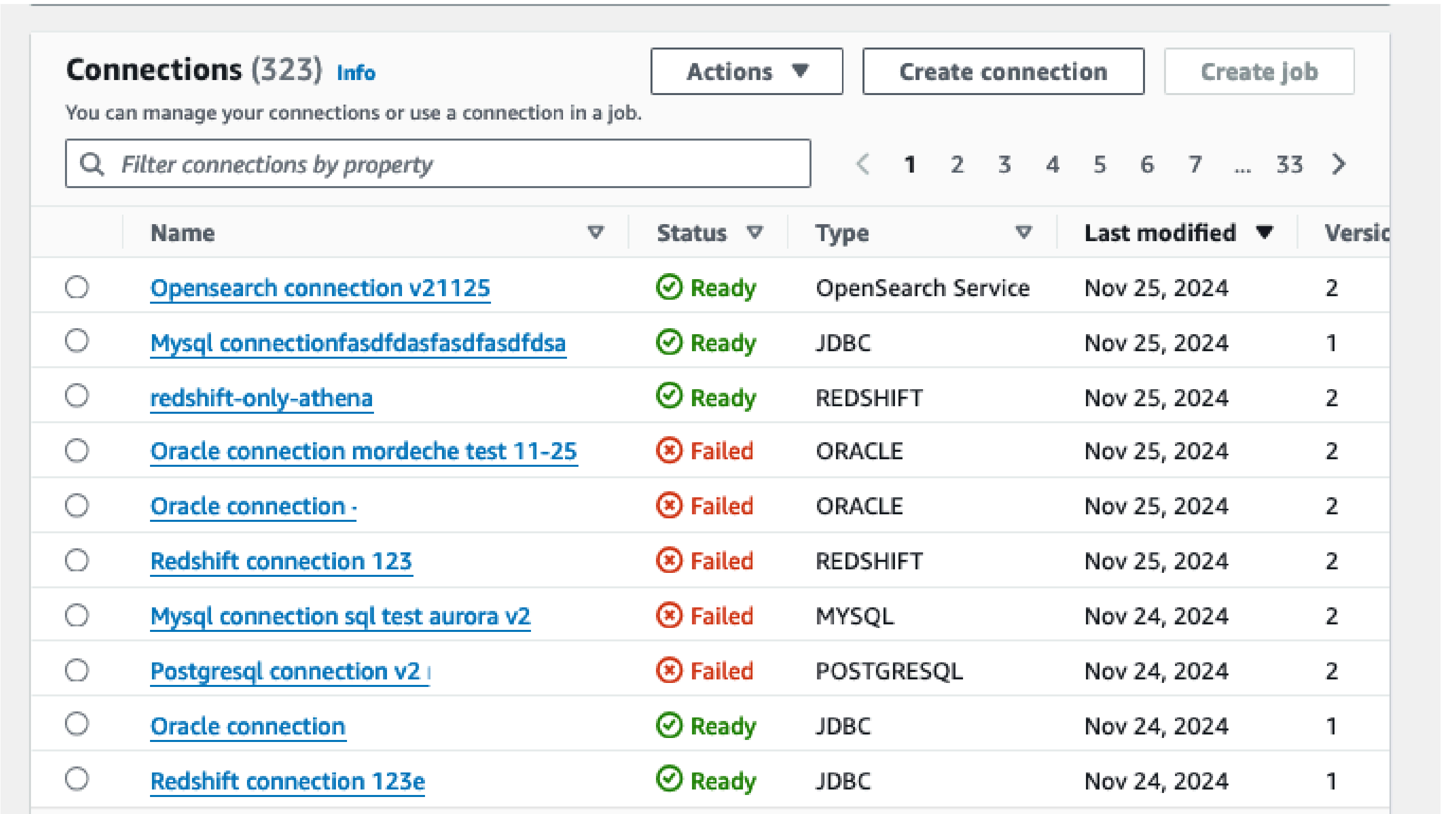
Version
2

Database
test

Host name

Port
5432

The connection version is also visible when viewing all your Connections.

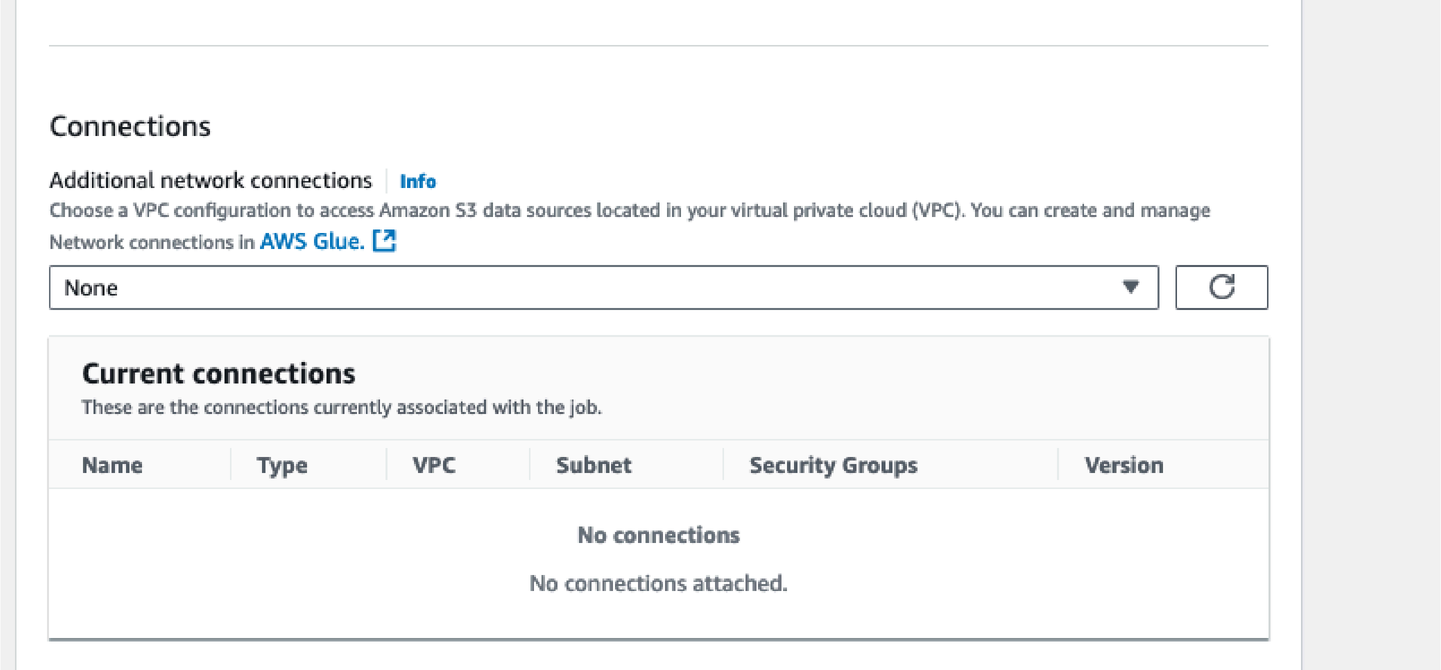


Connections (323) Info

You can manage your connections or use a connection in a job.

	Name	Status	Type	Last modified	Versic
○	Opensearch connection v21125	✓ Ready	OpenSearch Service	Nov 25, 2024	2
○	Mysql connectionfasdfdasfasdfasdfdsa	✓ Ready	JDBC	Nov 25, 2024	1
○	redshift-only-athena	✓ Ready	REDSHIFT	Nov 25, 2024	2
○	Oracle connection mordeche test 11-25	⊗ Failed	ORACLE	Nov 25, 2024	2
○	Oracle connection -	⊗ Failed	ORACLE	Nov 25, 2024	2
○	Redshift connection 123	⊗ Failed	REDSHIFT	Nov 25, 2024	2
○	Mysql connection sql test aurora v2	⊗ Failed	MYSQL	Nov 24, 2024	2
○	Postgresql connection v2	⊗ Failed	POSTGRESQL	Nov 24, 2024	2
○	Oracle connection	✓ Ready	JDBC	Nov 24, 2024	1
○	Redshift connection 123e	✓ Ready	JDBC	Nov 24, 2024	1

Finally, connection version is visible in the Job details tab for a job.



Connections

Additional network connections | Info

Choose a VPC configuration to access Amazon S3 data sources located in your virtual private cloud (VPC). You can create and manage Network connections in AWS Glue.

None ▼

Current connections

These are the connections currently associated with the job.

Name	Type	VPC	Subnet	Security Groups	Version

No connections

No connections attached.

With version 2 connections, you have the following expanded data connectivity capabilities:

- **Connection type discovery**: Support for creating connections using standardized templates. AWS Glue automatically discovers the connection types accessible by you and the required and optional inputs for a given connection type.
- **Reusability**: Connection definitions that are reusable across AWS data processing engines and tools like AWS Glue, Amazon Athena, and Amazon SageMaker AI. Connections now contain AthenaProperties, SparkProperties, PythonProperties which allow to specify compute environment/service specific connection properties in addition to the common properties stored in ConnectionProperties. Athena now creates Connections in AWS Glue by specifying Athena specific properties in the AthenaProperties property map.
- **Data preview**: Ability to browse metadata and preview data from connected sources.
- **Connector metadata**: Reusable connections may be used in order to discover table metadata.
- **Service linked secrets**: Users may provide necessary OAuth, basic or custom authentication credentials in the `CreateConnection` request. The CreateConnection API creates a Service Linked Secret in your account and stores the credentials on your behalf.

Supported authentication types

Unified connections supports the following authentication types:

- **BASIC –** Most database connection types and existing AWS Glue connection types support basic authentication, which is a username and password. Previously, the naming of the keys in SecretsManager were connector specific and, for example, may have been user, username, userName, opensearch.net.http.auth.user, etc. This is where unified connections standardized basic authentication connection types on USERNAME and PASSWORD keys.
- **OAUTH2 –** The majority of newly launched SaaS connection types support OAuth2 protocol.
- **CUSTOM –** A few connection types have some other authentication mechanism such as Google BigQuery where users are expected to provide the JSON which they get from Google BigQuery.

Considerations

When you create a unified connection for data sources, consider the following differences:

- When creating a unified connection via AWS Glue Studio, user credentials are stored in AWS Secrets Manager instead of the connection itself. This means jobs now need access to Secrets Manager.

- If jobs run in a VPC, they require either a VPC endpoint or NAT gateway to access AWS Secrets Manager and Secure Token Service (STS), which incurs additional costs.
- For certain data sources (Redshift, SQL Server, MySQL, Oracle, PostgreSQL), creating a unified connection via AWS Glue Studio requires access to AWS STS and AWS Secrets Manager. This is necessary to establish a secure connection and retrieve the required credentials for accessing these data sources within your Virtual Private Cloud (VPC).
- Creating a unified connection via AWS Glue Studio requires an IAM role with permissions to access AWS Secrets Manager and manage VPC resources (if using a VPC):
  - secretsmanager:GetSecretValue
  - secretsmanager:PutSecretValue
  - secretsmanager:DescribeSecret
  - ec2:CreateNetworkInterface
  - ec2:DeleteNetworkInterface
  - ec2:DescribeNetworkInterfaces
  - ec2:DescribeSubnets

Available connections

AWS Glue supports the following connection types:

- Adobe Analytics
- Adobe Marketo Engage
- Amazon Aurora (supported if the native JDBC driver is being used. Not all driver features can be leveraged)
- Amazon DocumentDB
- Amazon DynamoDB
- Amazon OpenSearch Service, for use with AWS Glue for Spark.
- Amazon Redshift
- Asana
- Azure Cosmos, for use of Azure Cosmos DB for NoSQL with AWS Glue ETL jobs
- Azure SQL, for use with AWS Glue for Spark.
- Blackbaud

- CircleCI
- Datadog
- Docusign Monitor
- Domo
- Dynatrace
- Facebook Ads
- Facebook Page Insights
- Freshdesk
- Freshsales
- Google Ads
- Google Analytics 4
- Google BigQuery, for use with AWS Glue for Spark.
- Google Search Console
- Google Sheets
- HubSpot
- Instagram Ads
- Intercom
- JDBC
- Jira Cloud
- Kafka
- Kustomer
- LinkedIn
- Mailchimp
- Microsoft Dynamics 365 CRM
- Microsoft Teams
- Mixpanel
- Monday
- MongoDB
- MongoDB Atlas

- Okta
- Oracle NetSuite
- Paypal
- Pendo
- Pipedrive
- Productboard
- QuickBooks
- Salesforce
- Salesforce Commerce Cloud
- Salesforce Marketing Cloud
- Salesforce Marketing Cloud Account Engagement (previously Salesforce Pardot)
- SAP HANA, for use with AWS Glue for Spark.
- SAP OData
- SendGrid
- ServiceNow
- Slack
- Smartsheet
- Snapchat Ads
- Stripe
- Snowflake, for use with AWS Glue for Spark.
- Teradata Vantage, when using AWS Glue for Spark.
- Twilio
- Vertica, for use with AWS Glue for Spark.
- WooCommerce
- Zendesk
- Zoho CRM
- Zoom Meetings
- Various Amazon Relational Database Service (Amazon RDS) offerings.
- Network (designates a connection to a data source that is in an Amazon Virtual Private Cloud (Amazon VPC))

With AWS Glue Studio, you can also create a connection for a *connector*. A connector is an optional code package that assists with accessing data stores in AWS Glue Studio. For more information, see Using connectors and connections with AWS Glue Studio

For information about how to connect to on-premises databases, see How to access and analyze on-premises data stores using AWS Glue at the AWS Big Data Blog website.

To create a connection with VPC configuration while using a custom IAM role, it must have the following VPC access actions:

- secretsmanager:GetSecretValue
- secretsmanager:PutSecretValue
- secretsmanager:DescribeSecret
- ec2:CreateNetworkInterface
- ec2:DeleteNetworkInterface
- ec2:DescribeNetworkInterfaces
- ec2:DescribeSubnets

Limitations

- You can't edit connections through the AWS Glue console if you created a v2 connection using AWS Glue APIs:
  - Amazon DocumentDB
  - Amazon Aurora
  - MariaDB
  - MongoDB Atlas
  - MongoDB

AWS Glue connection properties

This topic includes information about properties for AWS Glue connections.

Topics

- Required connection properties

- AWS Glue JDBC connection properties
- AWS Glue MongoDB and MongoDB Atlas connection properties
- Salesforce connection properties
- Snowflake connection
- Vertica connection
- SAP HANA connection
- Azure SQL connection
- Teradata Vantage connection
- OpenSearch Service connection
- Azure Cosmos connection
- AWS Glue SSL connection properties
- Apache Kafka connection properties for client authentication
- Google BigQuery connection
- Vertica connection

Required connection properties

When you define a connection on the AWS Glue console, you must provide values for the following properties:

Connection name

Enter a unique name for your connection.

Connection type

Choose **JDBC** or one of the specific connection types.

For details about the JDBC connection type, see the section called "JDBC connection properties"

Choose **Network** to connect to a data source within an Amazon Virtual Private Cloud environment (Amazon VPC)).

Depending on the type that you choose, the AWS Glue console displays other required fields. For example, if you choose **Amazon RDS**, you must then choose the database engine.

Require SSL connection

When you select this option, AWS Glue must verify that the connection to the data store is connected over a trusted Secure Sockets Layer (SSL).

For more information, including additional options that are available when you select this option, see the section called "SSL connection properties".

Select MSK cluster (Amazon managed streaming for Apache Kafka (MSK) only)

Specifies an MSK cluster from another AWS account.

Kafka bootstrap server URLs (Kafka only)

Specifies a comma-separated list of bootstrap server URLs. Include the port number. For example: b-1.vpc-test-2.o4q88o.c6.kafka.us-east-1.amazonaws.com:9094, b-2.vpc-test-2.o4q88o.c6.kafka.us-east-1.amazonaws.com:9094, b-3.vpc-test-2.o4q88o.c6.kafka.us-east-1.amazonaws.com:9094

AWS Glue JDBC connection properties

AWS Glue Studio now creates unified connections for MySQL, Oracle, PostgresSQL, Redshift, and SQL Server data sources, which requires additional steps for accessing Secrets Manager and VPC resources, which may incur extra costs. You can access these connections in AWS Glue Studio by choosing the connection name for the respective connection.

For more information, see Considerations.

AWS Glue can connect to the following data stores through a JDBC connection:

- Amazon Redshift
- Amazon Aurora
- Microsoft SQL Server
- MySQL
- Oracle
- PostgreSQL
- Snowflake, when using AWS Glue crawlers.
- Aurora (supported if the native JDBC driver is being used. Not all driver features can be leveraged)

- Amazon RDS for MariaDB

Important

Currently, an ETL job can use JDBC connections within only one subnet. If you have multiple data stores in a job, they must be on the same subnet, or accessible from the subnet.
If you choose to bring in your own JDBC driver versions for AWS Glue crawlers, your crawlers will consume resources in AWS Glue jobs and Amazon S3 to ensure your provided drivers are run in your environment. The additional usage of resources will be reflected in your account. Additionally, providing your own JDBC driver does not mean that the crawler is able to leverage all of the driver's features. Drivers are limited to the properties described in Defining connections in the Data Catalog.

The following are additional properties for the JDBC connection type.

JDBC URL

Enter the URL for your JDBC data store. For most database engines, this field is in the following format. In this format, replace *`protocol`*, *`host`*, *`port`*, and *`db_name`* with your own information.

```
jdbc:protocol://host:port/db_name
```

Depending on the database engine, a different JDBC URL format might be required. This format can have slightly different use of the colon (:) and slash (/) or different keywords to specify databases.

For JDBC to connect to the data store, a `db_name` in the data store is required. The `db_name` is used to establish a network connection with the supplied `username` and `password`. When connected, AWS Glue can access other databases in the data store to run a crawler or run an ETL job.

The following JDBC URL examples show the syntax for several database engines.

- To connect to an Amazon Redshift cluster data store with a `dev` database:

```
jdbc:redshift://xxx.us-east-1.redshift.amazonaws.com:8192/dev
```

- To connect to an Amazon RDS for MySQL data store with an `employee` database:

  ```
  jdbc:mysql://xxx-cluster.cluster-xxx.us-east-1.rds.amazonaws.com:3306/employee
  ```
- To connect to an Amazon RDS for PostgreSQL data store with an `employee` database:

  ```
  jdbc:postgresql://xxx-cluster.cluster-xxx.us-east-1.rds.amazonaws.com:5432/employee
  ```
- To connect to an Amazon RDS for Oracle data store with an `employee` service name:

  ```
  jdbc:oracle:thin://@xxx-cluster.cluster-xxx.us-east-1.rds.amazonaws.com:1521/employee
  ```

  The syntax for Amazon RDS for Oracle can follow the following patterns. In these patterns, replace *`host`*, *`port`*, *`service_name`*, and *`SID`* with your own information.

  - `jdbc:oracle:thin://@host:port/service_name`
  - `jdbc:oracle:thin://@host:port:SID`
- To connect to an Amazon RDS for Microsoft SQL Server data store with an `employee` database:

  ```
  jdbc:sqlserver://xxx-cluster.cluster-xxx.us-east-1.rds.amazonaws.com:1433;databaseName=employee
  ```

  The syntax for Amazon RDS for SQL Server can follow the following patterns. In these patterns, replace *`server_name`*, *`port`*, and *`db_name`* with your own information.

  - `jdbc:sqlserver://server_name:port;database=db_name`
  - `jdbc:sqlserver://server_name:port;databaseName=db_name`
- To connect to an Amazon Aurora PostgreSQL instance of the `employee` database, specify the endpoint for the database instance, the port, and the database name:

  ```
  jdbc:postgresql://employee_instance_1.xxxxxxxxxxxx.us-east-2.rds.amazonaws.com:5432/employee
  ```
- To connect to an Amazon RDS for MariaDB data store with an `employee` database, specify the endpoint for the database instance, the port, and the database name:

  ```
  jdbc:mysql://xxx-cluster.cluster-xxx.aws-region.rds.amazonaws.com:3306/employee
  ```

- **Warning**

  Snowflake JDBC connections are supported only by AWS Glue crawlers. When using the Snowflake connector in AWS Glue jobs, use the Snowflake connection type.

  To connect to a Snowflake instance of the `sample` database, specify the endpoint for the snowflake instance, the user, the database name, and the role name. You can optionally add the `warehouse` parameter.

  ```
  jdbc:snowflake://account_name.snowflakecomputing.com/?user=user_name&db=sample&role=role_name&warehouse=warehouse_name
  ```

  Important

  For Snowflake connections over JDBC, the order of parameters in the URL is enforced and must be ordered as `user`, db, `role_name`, and `warehouse`.

- To connect to a Snowflake instance of the `sample` database with AWS private link, specify the snowflake JDBC URL as follows:

  ```
  jdbc:snowflake://account_name.region.privatelink.snowflakecomputing.com/?user=user_name&db=sample&role=role_name&warehouse=warehouse_name
  ```

Username

Note

We recommend that you use an AWS secret to store connection credentials instead of supplying your user name and password directly. For more information, see Storing connection credentials in AWS Secrets Manager.

Provide a user name that has permission to access the JDBC data store.

Password

Enter the password for the user name that has access permission to the JDBC data store.

Port

Enter the port used in the JDBC URL to connect to an Amazon RDS Oracle instance. This field is only shown when **Require SSL connection** is selected for an Amazon RDS Oracle instance.

VPC

Choose the name of the virtual private cloud (VPC) that contains your data store. The AWS Glue console lists all VPCs for the current Region.

> ⚠ **Important**
>
> When working over a JDBC connection which is hosted off of AWS, such as with data from Snowflake, your VPC should have a NAT gateway which splits traffic into public and private subnets. The public subnet is used for connection to the external source, and the internal subnet is used for processing by AWS Glue. For information on configuring your Amazon VPC for external connections, read Connect to the internet or other networks using NAT devices and Setting up Amazon VPC for JDBC connections to Amazon RDS data stores from AWS Glue.

Subnet

Choose the subnet within the VPC that contains your data store. The AWS Glue console lists all subnets for the data store in your VPC.

Security groups

Choose the security groups that are associated with your data store. AWS Glue requires one or more security groups with an inbound source rule that allows AWS Glue to connect. The AWS Glue console lists all security groups that are granted inbound access to your VPC. AWS Glue associates these security groups with the elastic network interface that is attached to your VPC subnet.

JDBC Driver Class name - optional

Provide the custom JDBC driver class name:

- Postgres – org.postgresql.Driver
- MySQL – com.mysql.jdbc.Driver, com.mysql.cj.jdbc.Driver

- Redshift – com.amazon.redshift.jdbc.Driver, com.amazon.redshift.jdbc42.Driver
- Oracle – oracle.jdbc.driver.OracleDriver
- SQL Server – com.microsoft.sqlserver.jdbc.SQLServerDriver

JDBC Driver S3 Path - optional

Provide the Amazon S3 location to the custom JDBC driver. This is an absolute path to a .jar file. If you want to provide your own JDBC drivers to connect to your data souces for your crawler-supported databases,
you can specify values for parameters
`customJdbcDriverS3Path` and `customJdbcDriverClassName`.
Using a JDBC driver supplied by a customer is limited to the required Required connection properties.

AWS Glue MongoDB and MongoDB Atlas connection properties

The following are additional properties for the MongoDB or MongoDB Atlas connection type.

MongoDB URL

Enter the URL for your MongoDB or MongoDB Atlas data store:

- For MongoDB: mongodb://host:port/database. The host can be a hostname, IP address, or UNIX domain socket. If the connection string doesn't specify a port, it uses the default MongoDB port, 27017.
- For MongoDB Atlas: mongodb+srv://server.example.com/database. The host can be a hostname that follows corresponds to a DNS SRV record. The SRV format does not require a port and will use the default MongoDB port, 27017.

Username

> **Note**
>
> We recommend that you use an AWS secret to store connection credentials instead of supplying your user name and password directly. For more information, see Storing connection credentials in AWS Secrets Manager.

Provide a user name that has permission to access the JDBC data store.

Password

Enter the password for the user name that has access permission to the MongoDB or MongoDB Atlas data store.

Salesforce connection properties

The following are additional properties for the Salesforce connection type.

- `ENTITY_NAME`(String) - (Required) Used for Read/Write. The name of your Object in Salesforce.
- `API_VERSION`(String) - (Required) Used for Read/Write. Salesforce Rest API version you want to use.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.
- `IMPORT_DELETED_RECORDS`(String) - Default: FALSE. Used for read. To get the delete records while querying.
- `WRITE_OPERATION`(String) - Default: INSERT. Used for write. Value should be INSERT, UPDATE, UPSERT, DELETE.
- `ID_FIELD_NAMES`(String) - Default : null. Used only for UPSERT.

Snowflake connection

The following properties are used to set up a Snowflake connection used in AWS Glue ETL jobs. When crawling Snowflake, use a JDBC connection.

Snowflake URL

The URL of your Snowflake endpoint. For more information about Snowflake endpoint URLs, see Connecting to Your Accounts in the Snowflake documentation.

AWS Secret

The **Secret name** of a secret in AWS Secrets Manager. AWS Glue will connect to Snowflake using the `sfUser` and `sfPassword` keys of your secret.

Snowflake role (optional)

A Snowflake security role AWS Glue will use when connecting.

Use the following properties when configuring a connection to a Snowflake endpoint hosted in Amazon VPC using AWS PrivateLink.

VPC

Choose the name of the virtual private cloud (VPC) that contains your data store. The AWS Glue console lists all VPCs for the current Region.

Subnet

Choose the subnet within the VPC that contains your data store. The AWS Glue console lists all subnets for the data store in your VPC.

Security groups

Choose the security groups that are associated with your data store. AWS Glue requires one or more security groups with an inbound source rule that allows AWS Glue to connect. The AWS Glue console lists all security groups that are granted inbound access to your VPC. AWS Glue associates these security groups with the elastic network interface that is attached to your VPC subnet.

Vertica connection

Use the following properties to set up a Vertica connection for AWS Glue ETL jobs.

Vertica Host

The hostname of your Vertica installation.

Vertica Port

The port your Vertica installation is available through.

AWS Secret

The **Secret name** of a secret in AWS Secrets Manager. AWS Glue will connect to Vertica using the keys of your secret.

Use the following properties when configuring a connection to a Vertica endpoint hosted in Amazon VPC.

VPC

Choose the name of the virtual private cloud (VPC) that contains your data store. The AWS Glue console lists all VPCs for the current Region.

Subnet

Choose the subnet within the VPC that contains your data store. The AWS Glue console lists all subnets for the data store in your VPC.

Security groups

Choose the security groups that are associated with your data store. AWS Glue requires one or more security groups with an inbound source rule that allows AWS Glue to connect. The AWS Glue console lists all security groups that are granted inbound access to your VPC. AWS Glue associates these security groups with the elastic network interface that is attached to your VPC subnet.

SAP HANA connection

Use the following properties to set up a SAP HANA connection for AWS Glue ETL jobs.

SAP HANA URL

A SAP JDBC URL.

SAP HANA JDBC URLs are in the form `jdbc:sap://saphanaHostname:saphanaPort/?databaseName=saphanaDBname,ParameterNa`

AWS Glue requires the following JDBC URL parameters:

- `databaseName` – A default database in SAP HANA to connect to.

AWS Secret

The **Secret name** of a secret in AWS Secrets Manager. AWS Glue will connect to SAP HANA using the keys of your secret.

Use the following properties when configuring a connection to a SAP HANA endpoint hosted in Amazon VPC:

VPC

Choose the name of the virtual private cloud (VPC) that contains your data store. The AWS Glue console lists all VPCs for the current Region.

Subnet

Choose the subnet within the VPC that contains your data store. The AWS Glue console lists all subnets for the data store in your VPC.

Security groups

Choose the security groups that are associated with your data store. AWS Glue requires one or more security groups with an inbound source rule that allows AWS Glue to connect. The AWS Glue console lists all security groups that are granted inbound access to your VPC. AWS Glue associates these security groups with the elastic network interface that is attached to your VPC subnet.

Azure SQL connection

Use the following properties to set up a Azure SQL connection for AWS Glue ETL jobs.

Azure SQL URL

The JDBC URL of an Azure SQL endpoint.

The URL must be in the following format: `jdbc:sqlserver://databaseServerName:databasePort;databaseName=azuresqlDBname;.`

AWS Glue requires the following URL properties:

- `databaseName` – A default database in Azure SQL to connect to.

For more information about JDBC URLs for Azure SQL Managed Instances, see the Microsoft documentation.

AWS Secret

The **Secret name** of a secret in AWS Secrets Manager. AWS Glue will connect to Azure SQL using the keys of your secret.

Teradata Vantage connection

Use the following properties to set up a Teradata Vantage connection for AWS Glue ETL jobs.

Teradata URL

To connect to a Teradata instance specify the hostname for the database instance and relevant Teradata parameters:

```
jdbc:teradata://teradataHostname/ParameterName=ParameterValue,ParameterName=Pa
```

AWS Glue supports the following JDBC URL parameters:

- `DATABASE_NAME` – A default database in Teradata to connect to.
- `DBS_PORT` – Specifies the Teradata port, if nonstandard.

AWS Secret

The **Secret name** of a secret in AWS Secrets Manager. AWS Glue will connect to Teradata Vantage using the keys of your secret.

Use the following properties when configuring a connection to a Teradata Vantage endpoint hosted in Amazon VPC:

VPC

Choose the name of the virtual private cloud (VPC) that contains your data store. The AWS Glue console lists all VPCs for the current Region.

Subnet

Choose the subnet within the VPC that contains your data store. The AWS Glue console lists all subnets for the data store in your VPC.

Security groups

Choose the security groups that are associated with your data store. AWS Glue requires one or more security groups with an inbound source rule that allows AWS Glue to connect. The AWS

Glue console lists all security groups that are granted inbound access to your VPC. AWS Glue associates these security groups with the elastic network interface that is attached to your VPC subnet.

OpenSearch Service connection

Use the following properties to set up a OpenSearch Service connection for AWS Glue ETL jobs.

Domain endpoint

An Amazon OpenSearch Service domain endpoint will have the following default form, https://search-*`domainName`*-*`unstructuredIdContent`*.*`region`*.es.amazonaws.com. For more information on identifying your domain endpoint, see Creating and managing Amazon OpenSearch Service domains in the Amazon OpenSearch Service documentation.

Port

The port open on the endpoint.

AWS Secret

The **Secret name** of a secret in AWS Secrets Manager. AWS Glue will connect to OpenSearch Service using the keys of your secret.

Use the following properties when configuring a connection to a OpenSearch Service endpoint hosted in Amazon VPC:

VPC

Choose the name of the virtual private cloud (VPC) that contains your data store. The AWS Glue console lists all VPCs for the current Region.

Subnet

Choose the subnet within the VPC that contains your data store. The AWS Glue console lists all subnets for the data store in your VPC.

Security groups

Choose the security groups that are associated with your data store. AWS Glue requires one or more security groups with an inbound source rule that allows AWS Glue to connect. The AWS

Glue console lists all security groups that are granted inbound access to your VPC. AWS Glue associates these security groups with the elastic network interface that is attached to your VPC subnet.

Azure Cosmos connection

Use the following properties to set up a Azure Cosmos connection for AWS Glue ETL jobs.

Azure Cosmos DB Account Endpoint URI

The endpoint used to connect to Azure Cosmos. For more information, see the Azure documentation.

AWS Secret

The **Secret name** of a secret in AWS Secrets Manager. AWS Glue will connect to Azure Cosmos using the keys of your secret.

AWS Glue SSL connection properties

The following are details about the **Require SSL connection** property.

If you do not require SSL connection, AWS Glue ignores failures when it uses SSL to encrypt a connection to the data store. See the documentation for your data store for configuration instructions. When you select this option, the job run, crawler, or ETL statements in a development endpoint fail when AWS Glue cannot connect.

> **Note**
>
> Snowflake supports an SSL connection by default, so this property is not applicable for Snowflake.

This option is validated on the AWS Glue client side. For JDBC connections, AWS Glue only connects over SSL with certificate and host name validation. SSL connection support is available for:

- Oracle Database
- Microsoft SQL Server

- PostgreSQL
- Amazon Redshift
- MySQL (Amazon RDS instances only)
- Amazon Aurora MySQL (Amazon RDS instances only)
- Amazon Aurora PostgreSQL (Amazon RDS instances only)
- Kafka, which includes Amazon Managed Streaming for Apache Kafka
- MongoDB

> **Note**
>
> To enable an **Amazon RDS Oracle** data store to use **Require SSL connection**, you must create and attach an option group to the Oracle instance.
>
> 1. Sign in to the AWS Management Console and open the Amazon RDS console at https://console.aws.amazon.com/rds/.
> 2. Add an **Option group** to the Amazon RDS Oracle instance. For more information about how to add an option group on the Amazon RDS console, see Creating an Option Group
> 3. Add an **Option** to the option group for **SSL**. The **Port** you specify for SSL is later used when you create an AWS Glue JDBC connection URL for the Amazon RDS Oracle instance. For more information about how to add an option on the Amazon RDS console, see Adding an Option to an Option Group in the *Amazon RDS User Guide*. For more information about the Oracle SSL option, see Oracle SSL in the *Amazon RDS User Guide*.
> 4. On the AWS Glue console, create a connection to the Amazon RDS Oracle instance. In the connection definition, select **Require SSL connection**. When requested, enter the **Port** that you used in the Amazon RDS Oracle SSL option.

The following additional optional properties are available when **Require SSL connection** is selected for a connection:

Custom JDBC certificate in S3

If you have a certificate that you are currently using for SSL communication with your on-premises or cloud databases, you can use that certificate for SSL connections to AWS Glue data sources or targets. Enter an Amazon Simple Storage Service (Amazon S3) location that contains a custom root certificate. AWS Glue uses this certificate to establish an SSL connection to the database. AWS Glue handles only X.509 certificates. The certificate must be DER-encoded and supplied in base64 encoding PEM format.

If this field is left blank, the default certificate is used.

Custom JDBC certificate string

Enter certificate information specific to your JDBC database. This string is used for domain matching or distinguished name (DN) matching. For Oracle Database, this string maps to the `SSL_SERVER_CERT_DN` parameter in the security section of the `tnsnames.ora` file. For Microsoft SQL Server, this string is used as `hostNameInCertificate`.

The following is an example for the Oracle Database `SSL_SERVER_CERT_DN` parameter.

```
cn=sales,cn=OracleContext,dc=us,dc=example,dc=com
```

Kafka private CA certificate location

If you have a certificate that you are currently using for SSL communication with your Kafka data store, you can use that certificate with your AWS Glue connection. This option is required for Kafka data stores, and optional for Amazon Managed Streaming for Apache Kafka data stores. Enter an Amazon Simple Storage Service (Amazon S3) location that contains a custom root certificate. AWS Glue uses this certificate to establish an SSL connection to the Kafka data store. AWS Glue handles only X.509 certificates. The certificate must be DER-encoded and supplied in base64 encoding PEM format.

Skip certificate validation

Select the **Skip certificate validation** check box to skip validation of the custom certificate by AWS Glue. If you choose to validate, AWS Glue validates the signature algorithm and subject public key algorithm for the certificate. If the certificate fails validation, any ETL job or crawler that uses the connection fails.

The only permitted signature algorithms are SHA256withRSA, SHA384withRSA, or SHA512withRSA. For the subject public key algorithm, the key length must be at least 2048.

Kafka client keystore location

The Amazon S3 location of the client keystore file for Kafka client side authentication. Path must be in the form s3://bucket/prefix/filename.jks. It must end with the file name and .jks extension.

Kafka client keystore password (optional)

The password to access the provided keystore.

Kafka client key password (optional)

A keystore can consist of multiple keys, so this is the password to access the client key to be used with the Kafka server side key.

Apache Kafka connection properties for client authentication

AWS Glue supports the Simple Authentication and Security Layer (SASL) framework for authentication when you create an Apache Kafka connection. The SASL framework supports various mechanisms of authentication, and AWS Glue offers the SCRAM (user name and password), GSSAPI (Kerberos protocol), and PLAIN protocols.

Use AWS Glue Studio to configure one of the following client authentication methods. For more information, see Creating connections for connectors in the AWS Glue Studio user guide.

- None - No authentication. This is useful if creating a connection for testing purposes.
- SASL/SCRAM-SHA-512 - Choosing this authentication method will allow you to specify authentication credentials. There are two options available:
  - Use AWS Secrets Manager (recommended) - if you select this option, you can store your user name and password in AWS Secrets Manager and let AWS Glue access them when needed. Specify the secret that stores the SSL or SASL authentication credentials. For more information, see Storing connection credentials in AWS Secrets Manager.
  - Provide a user name and password directly.
- SASL/GSSAPI (Kerberos) - if you select this option, you can select the location of the keytab file, krb5.conf file and enter the Kerberos principal name and Kerberos service name. The locations for the keytab file and krb5.conf file must be in an Amazon S3 location. Since MSK does not yet support SASL/GSSAPI, this option is only available for customer managed Apache Kafka clusters. For more information, see MIT Kerberos Documentation: Keytab .
- SASL/PLAIN - choose this authentication method to specify authentication credentials. There are two options available:
  - Use AWS Secrets Manager (recommended) - if you select this option, you can store your credentials in AWS Secrets Manager and let AWS Glue access the information when needed. Specify the secret that stores the SSL or SASL authentication credentials.

  - Provide username and password directly.
- SSL Client Authentication - if you select this option, you can you can select the location of the Kafka client keystore by browsing Amazon S3. Optionally, you can enter the Kafka client keystore password and Kafka client key password.

Google BigQuery connection

The following properties are used to set up a Google BigQuery connection used in AWS Glue ETL jobs. For more information, see the section called "BigQuery connections".

AWS Secret

> The **Secret name** of a secret in AWS Secrets Manager. AWS Glue ETL jobs will connect to Google BigQuery using the `credentials` key of your secret.

Vertica connection

The following properties are used to set up a Vertica connection used in AWS Glue ETL jobs. For more information, see the section called "Vertica connections".

Storing connection credentials in AWS Secrets Manager

We recommend that you use AWS Secrets Manager to supply connection credentials for your data store. Using Secrets Manager this way lets AWS Glue access your secret at runtime for ETL jobs and crawler runs, and helps keep your credentials secure.

Prerequisites

To use Secrets Manager with AWS Glue, you must grant your IAM role for AWS Glue permission to retrieve secret values. The AWS managed policy `AWSGlueServiceRole` doesn't include AWS Secrets Manager permissions. For example IAM policies, see Example: Permission to retrieve secret values in the *AWS Secrets Manager User Guide.*

Depending on your network setup, you might also need to create a VPC endpoint to establish a private connection between your VPC and Secrets Manager. For more information, see Using an AWS Secrets Manager VPC endpoint.

To create a secret for AWS Glue

1. Follow the instructions in Create and manage secrets in the *AWS Secrets Manager User Guide*. The following example JSON shows how to specify your credentials in the **Plaintext** tab when you create a secret for AWS Glue.

```
{
  "username": "EXAMPLE-USERNAME",
  "password": "EXAMPLE-PASSWORD"
}
```

2. Associate your secret with a connection using the AWS Glue Studio interface. For detailed instructions, see Creating connections for connectors in the *AWS Glue Studio User Guide.*

Adding an AWS Glue connection

You can connect to data sources in AWS Glue for Spark programmatically. For more information, see Connection types and options for ETL in AWS Glue for Spark

You can also use the AWS Glue console to add, edit, delete, and test connections. For information about AWS Glue connections, see Connecting to data.

Topics

- Connecting to Adobe Analytics
- Connecting to Adobe Marketo Engage
- Connecting to Amazon Redshift in AWS Glue Studio
- Connecting to Asana
- Connecting to Azure Cosmos DB in AWS Glue Studio
- Connecting to Azure SQL in AWS Glue Studio
- Connecting to Blackbaud Raiser's Edge NXT
- Connecting to CircleCI
- Connecting to Datadog
- Connecting to Docusign Monitor
- Connecting to Domo
- Connecting to Dynatrace
- Connecting to Facebook Ads

- Connecting to Facebook Page Insights
- Connecting to Freshdesk
- Connecting to Freshsales
- Connecting to Google Ads
- Connecting to Google Analytics 4
- Connecting to Google BigQuery in AWS Glue Studio
- Connecting to Google Search Console
- Connecting to Google Sheets
- Connecting to HubSpot
- Connecting to Instagram Ads
- Connecting to Intercom in AWS Glue Studio
- Connecting to Jira Cloud
- Connecting to Kustomer
- Connecting to LinkedIn
- Connecting to Mailchimp
- Connecting to Microsoft Dynamics 365 CRM
- Connecting to Microsoft Teams
- Connecting to Mixpanel
- Connecting to Monday
- Connecting to MongoDB in AWS Glue Studio
- Connecting to Oracle NetSuite
- Connecting to OpenSearch Service in AWS Glue Studio
- Connecting to Okta
- Connecting to PayPal
- Connecting to Pendo
- Connecting to Pipedrive
- Connecting to Productboard
- Connecting to QuickBooks
- Connecting to Salesforce
- Connecting to Salesforce Marketing Cloud

- Connecting to Salesforce Commerce Cloud
- Connecting to Salesforce Marketing Cloud Account Engagement
- Connecting to SAP HANA in AWS Glue Studio
- Connecting to SAP OData
- Connecting to SendGrid
- Connecting to ServiceNow
- Connecting to Slack in AWS Glue Studio
- Connecting to Smartsheet
- Connecting to Snapchat Ads in AWS Glue Studio
- Connecting to Snowflake in AWS Glue Studio
- Connecting to Stripe in AWS Glue Studio
- Connecting to Teradata Vantage in AWS Glue Studio
- Connecting to Twilio
- Connecting to Vertica in AWS Glue Studio
- Connecting to WooCommerce
- Connecting to Zendesk
- Connecting to Zoho CRM
- Connecting to Zoom Meetings
- Adding a JDBC connection using your own JDBC drivers

Connecting to Adobe Analytics

Adobe Analytics is a robust data analysis platform that collects data from multi-channel digital experiences that support the customer journey and provides tools for analyzing the data. It’s a platform commonly used by marketers and business analysts for business analytics purposes. If you're a Adobe Analytics user, you can connect AWS Glue to your Adobe Analytics account. Then, you can use Adobe Analytics as a data source in your ETL Jobs. Run these jobs to transfer data between Adobe Analytics and AWS services or other supported applications.

Topics

- AWS Glue support for Adobe Analytics
- Policies containing the API operations for creating and using connections

- Configuring Adobe Analytics
- Configuring Adobe Analytics connections
- Reading from Adobe Analytics entities
- Adobe Analytics connection options
- Creating an Adobe Analytics account
- Limitations

AWS Glue support for Adobe Analytics

AWS Glue supports Adobe Analytics as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Adobe Analytics.

Supported as a target?

No.

Supported Adobe Analytics API versions

v2.0

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
                "glue:DescribeEntity"
            ],
            "Resource": "*"
        }
    ]
}
```

If you don't want to use the preceding method, alternatively, use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Adobe Analytics

Before you can use AWS Glue to transfer from Adobe Analytics, you must meet the following requirements:

Minimum requirements

- You have an Adobe Analytics account with email and password. For more information about creating an account, see Creating a Adobe Analytics account.
- Your Adobe Analytics account is enabled for API access. API access is enabled by default for the Select, Prime, and Ultimate editions.

If you meet these requirements, you’re ready to connect AWS Glue to your Adobe Analytics account. For typical connections, you don't need do anything else in Adobe Analytics.

Configuring Adobe Analytics connections

Adobe Analytics supports `AUTHORIZATION_CODE` grant type for `OAuth2`.

This grant type is considered “three-legged” `OAuth` as it relies on redirecting users to the third-party authorization server to authenticate the user. Users may opt to create their own connected app in Adobe Analytics and provide their own client ID and client secret when creating connections through the AWS Glue console. In this scenario, they will still be redirected to Adobe Analytics to login and authorize AWS Glue to access their resources.

This grant type results in a refresh token and access token. The access token is short lived, and may be refreshed automatically without user interaction using the refresh token.

For public Adobe Analytics documentation on creating a connected app for AUTHORIZATION_CODE OAuth flow, see Adobe Analytics APIs .

To configure a Adobe Analytics connection:

1. In AWS Secrets Manager, create a secret with the following details:

   For customer managed connected app – Secret should contain the connected app Consumer Secret with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as key.

   > **Note**
   >
   > It is a must to create a secret per connection in AWS Glue.

2. In AWS Glue Studio, create a connection under **Data Connections** by following the steps below:
   1. When selecting a **Connection type**, select Adobe Analytics.
   2. Provide the `x_api_key, instanceUrl` of the Adobe Analytics you want to connect to.
   3. Select the IAM role for which AWS Glue can assume and has permissions for following actions:

      JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
          "ec2:DeleteNetworkInterface"
        ],
        "Resource": "*"
      }
    ]
  }
```

   d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.
   e. Select the network options if you want to use your network.
3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Adobe Analytics entities

Prerequisites

An Adobe Analytics Object you would like to read from. Refer the supported entities table below to check the available entities.

Supported entities

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Annotation	Yes	Yes	Yes	Yes	No
Calculated Metrics	Yes	Yes	Yes	Yes	No
Calculated Metrics Function	Yes	No	No	Yes	No
Component Metadata Shares	Yes	Yes	No	Yes	No
Date Ranges	Yes	Yes	No	Yes	No
Dimensions	Yes	No	No	Yes	No

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Metrics	Yes	No	No	Yes	No
Projects	Yes	No	No	Yes	No
Reports Top Item	Yes	Yes	No	Yes	No
Segments	Yes	Yes	Yes	Yes	No
Usage Logs	Yes	Yes	No	Yes	No

Example

```
adobeAnalytics_read = glueContext.create_dynamic_frame.from_options(
     connection_type="adobeanalytics",
     connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "annotation/ex*****",
        "API_VERSION": "v2.0"
 })
```

Adobe Analytics entity and field details

- Annotations
- Calculated Metrics
- Component Meta Data
- Date Ranges
- Dimensions
- Metrics
- Projects
- Reports
- Segments
- Users
- Usage Logs

Adobe Analytics connection options

The following are connection options for Adobe Analytics:

- `ENTITY_NAME`(String) – (Required) Used for Read/Write. The name of your Object in Adobe Analytics.
- `API_VERSION`(String) – (Required) Used for Read/Write. Adobe Analytics Rest API version you want to use. For example: v2.0.
- `X_API_KEY`(String) – (Required) Used for Read/Write. It is required to authenticate the developer or application making requests to the API.
- `SELECTED_FIELDS`(List<String>) – Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) – Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) – Default: empty. Used for Read. Full Spark SQL query.

Creating an Adobe Analytics account

1. Register for the exchange partner program, by accessing Adobe Partner program.
2. Choose **Join Exchange Program**.
3. Register or create an account using your corporate email address.
4. From the suggestion box, select the appropriate company that has an Adobe Analytics product subscription.
5. Ensure that the account gets registered with a valid organization (from the list available) that has an active Adobe Analytics subscription.
6. After the company administration's approval, activate your account by clicking on the link in the approval email.

Verifying if the account that you created has access to Adobe Analytics service

1. Log in to Adobe Admin Console.
2. Check the organization name at the top-right corner of the page to ensure that you have logged in to the correct company.
3. Select **Products** and verify if Adobe Analytics is available.

Note

If no organization is available, or Adobe Analytics product is greyed out or is unavailable, it is likely that your account is not associated with an organization and/or has no active Adobe Analytics subscription. Contact your system admin to request access for this service for your account.

Creating a project and `OAuth2.0` credentials

1. Log in to Adobe Analytics account where you want the OAuth 2.0 app to be created.
2. Select **Project** and then **Create a new project**.
3. To add a project, select **Add to project**, and then select **API**.
4. Select **Adobe Analytics API**.
5. Select **OAUTH** as user authentication.
6. Select **Web** as `OAUTH` and provide the redirect URI.

   For redirect URI and its pattern, see the following:

   - `OAuth 2.0` default redirect URI – A default redirect URI is the URL of the page that Adobe will access during the authentication process. For example, `https://ap-southeast-2.console.aws.amazon.com/appflow/oauth`
   - OAuth 2.0 Redirect URI pattern – A Redirect URI pattern is a URI path (or comma-separated list of paths) to which Adobe can redirect (if requested) when the login flow is complete. For example, `https://ap-southeast-2\\.console\\.aws\\.amazon\\.com`
7. Add the following scopes:
   - `openid`
   - `read_organizations`
   - `additional_info.projectedProductContext`
   - `additional_info.job_function`
8. Choose **Save credential**.
9. After the app is created, copy the `Client ID` and `Client Secret` values to a text file.

Limitations

The following are limitations for the Adobe Analytics connector:

- Adobe Analytics doesn't support field based and record-based partitioning. Field based partitioning is not supported as you cannot query fields that you partition. Record based partitioning cannot be supported as there is no provision to get 'offset' for pagination.
- In the `Report Top Item` entity, the `startDate` and `endDate` query parameters are not functioning as expected. The response is not being filtered based on these parameters, which is causing issues with the filter and incremental flow for this entity.
- For the `Annotation`, `Calculated Metrics`, `Calculated Metrics Function`, `Date Ranges`, `Dimension`, `Metric`, `Project`, `Report Top Items`, and `Segment` entities, the `locale` query parameter specifies which language is to be used for localized sections of responses and does not filter the records. For example, `locale="ja_JP"` will display the data in Japanese.
- `Report Top Item` entity – filter on `dateRange` and `lookupNoneValues` fields are currently not working.
- `Segment` entity: with filter value `includeType="templates"`, filters on other fields are not working.
- `Date Range` entity – filter on `curatedRsid` field is not working.
- `Metric entity` entity – filter on segmentable field with "false" value gives result for both true and false value.

Connecting to Adobe Marketo Engage

Adobe Marketo Engage is a marketing automation platform that enables marketers to manage personalized multi-channel programs and campaigns to prospects and customers.

Topics

- AWS Glue support for Adobe Marketo Engage
- Policies containing the API operations for creating and using connections
- Configuring Adobe Marketo Engage
- Configuring Adobe Marketo Engage connections
- Reading from Adobe Marketo Engage entities
- Adobe Marketo Engage connection options

- Limitations and notes for Adobe Marketo Engage connector

AWS Glue support for Adobe Marketo Engage

AWS Glue supports Adobe Marketo Engage as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Adobe Marketo Engage.

Supported as a target?

No.

Supported Adobe Marketo Engage API versions

The following Adobe Marketo Engage API versions are supported:

- v1

For entity support per version specific, see Supported entities for Source.

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
```

```
        }
    ]
}
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Adobe Marketo Engage

Before you can use AWS Glue to transfer data from Adobe Marketo Engage, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- You have a Adobe Marketo Engage account with client credentials.
- Your Adobe Marketo Engage account has API access with a valid license.

If you meet these requirements, you’re ready to connect AWS Glue to your Adobe Marketo Engage account. For typical connections, you don't need do anything else in Adobe Marketo Engage.

Getting OAuth 2.0 credentials

To obtain API credentials so that you can make authenticated calls to your instance, see REST API in the Adobe Marketo Engage Developer Guide.

Configuring Adobe Marketo Engage connections

Adobe Marketo Engage supports the CLIENT CREDENTIALS grant type for OAuth2.

- This grant type is considered 2-legged OAuth 2.0 as it is used by clients to obtain an access token outside of the context of a user. AWS Glue is able to use the client ID and client secret to authenticate the Adobe Marketo Engage APIs which are provided by custom services that you define.
- Each custom service is owned by an API-only user which has a set of roles and permissions which authorize the service to perform specific actions. An access token is associated with a single custom service.
- This grant type results in an access token which is short lived, and may be renewed by calling an identity endpoint.
- For public Adobe Marketo Engage documentation for OAuth 2.0 with client credentials, see Authentication in the Adobe Marketo Engage Developer Guide.

To configure a Adobe Marketo Engage connection:

1. In AWS Secrets Manager, create a secret with the following details:
   a. For the customer managed connected app, the Secret should contain the connected app Consumer Secret with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as key.
   b. Note: You must create a secret per connection in AWS Glue.
2. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Connection type**, select Adobe Marketo Engage.
   b. Provide the `INSTANCE_URL` of the Adobe Marketo Engage instance you want to connect to.
   c. Select the AWS IAM role which AWS Glue can assume and has permissions for the following actions:

      JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
          "ec2:DescribeNetworkInterfaces",
          "ec2:DeleteNetworkInterface"
        ],
        "Resource": "*"
      }
    ]
  }
```

   d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.
   e. Select the network options if you want to use your network.

3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Adobe Marketo Engage entities

Prerequisite

A Adobe Marketo Engage object you would like to read from. You will need the object name such as leads or activities or customobjects. The following tables shows the supported entities.

Supported entities for source (synchronous):

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
leads	Yes	Yes	No	Yes	No
activities	Yes	Yes	No	Yes	No
customobjects	Yes	Yes	No	Yes	No

Supported entities for source (asynchronous):

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
leads	Yes	No	No	Yes	Yes

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
activities	Yes	No	No	Yes	No
customobjects	Yes	No	No	Yes	Yes

Example:

```
adobe-marketo-engage_read = glueContext.create_dynamic_frame.from_options(
    connection_type="adobe-marketo-engage",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "leads",
        "API_VERSION": "v2",
        "INSTANCE_URL": "https://539-t**-6**.mktorest.com"
    }
```

Adobe Marketo Engage entity and field details:

Entities with static metadata:

Entity	Field	Data type	Supported operators
activities	sinceDatetime (only supported in synchronous)	DateTime	>= (only for synchronous mode)
	createdAt (only supported in asynchronous)	DateTime	between (only for asynchronous mode)
	activitiesTypeId	Integer	=
	adobe-marketo-engageGUID	Long	= (only for synchronous mode)
	leadId	Long	N/A

Entity	Field	Data type	Supported operators
	activityDate	DateTime	N/A
	campaignId	Long	N/A
	primaryAttributeValueId	Integer	N/A
	primaryAttributeValue	String	N/A
	attributes	String	N/A

Entities with dynamic metadata:

For the following entities, Adobe Marketo Engage provides endpoints to fetch metadata dynamically, so that operator support is captured at the datatype level for each entity.

Entity	Data type	Supported operators
leads	Integer	= (only for synchronous mode)
	DateTime	between (only for asynchronous mode)
	String	= (only for synchronous mode)
	Long	N/A
	Boolean	N/A
	Date	N/A
	Float	N/A
customobjects	Integer	N/A

Entity	Data type	Supported operators
	DateTime	between (only for asynchronous mode)
	String	= (only for synchronous mode)
	Date	N/A
	Long	N/A
	Boolean	N/A
	Float	N/A

Partitioning queries

You can provide the additional Spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, and `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition the query.
- `LOWER_BOUND`: an **inclusive** lower bound value of the chosen partition field.

  For the DateTime field, we accept the value in ISO format.

  Example of valid value:

  ```
  "2024-07-01T00:00:00.000Z"
  ```
- `UPPER_BOUND`: an **exclusive** upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: the number of partitions.

The following table describes the entity partitioning field support details:

Entity name	Partitioning fields	Data type
leads	createdAt	DateTime
	updateAt	DateTime
customobjects	updatedAt	DateTime

Example:

```
adobe-marketo-engage_read = glueContext.create_dynamic_frame.from_options(
    connection_type="adobe-marketo-engage",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "leads",
        "API_VERSION": "v1",
        "PARTITION_FIELD": "createdAt"
        "LOWER_BOUND": "2024-07-01T00:00:00.000Z"
        "UPPER_BOUND": "2024-07-02T00:00:00.000Z"
        "NUM_PARTITIONS": "10"
    }
```

Adobe Marketo Engage connection options

The following are connection options for Adobe Marketo Engage:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your object in Adobe Marketo Engage.
- `API_VERSION`(String) - (Required) Used for Read. Adobe Marketo Engage Rest API version you want to use. For example: v1.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.

- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.
- `TRANSFER_MODE`(String) - Default: SYNC. Used for asynchronous read.

Limitations and notes for Adobe Marketo Engage connector

The following are limitations or notes for the Adobe Marketo Engage connector:

- 'sinceDatetime' and 'activityTypeId' are mandatory filter parameters for the Sync Activities entity.
- Subscriptions are allocated 50,000 API calls per day (which resets daily at 12:00AM CST). Additional daily capacity may be purchased as part of a Adobe Marketo Engage subscription.
- The maximum time span for the date range filter (`createdAt` or `updatedAt`) is 31 days.
- Subscriptions are allocated a maximum of 10 bulk extract jobs in the queue at any given time.
- By default, extract jobs are limited to 500 MB per day (which resets daily at 12:00AM CST). Additional daily capacity may be purchased as part of a Adobe Marketo Engage subscription.
- The maximum number of concurrent export jobs is 2.
- The maximum number of queued export jobs (inclusive of currently exporting jobs) is 10.
- The maximum allowed file size is 1 GB to extract from a Bulk job.
- Once an asynchronous job is created, the file retention period will be 7 days before expiring.
- `createdAt` or `updatedAt` are mandatory filter parameters for the Async Leads entity.
- `createdAt` is a mandatory filter parameter for the Async Activities entity.
- `updatedAt` is a mandatory filter parameter for the Async Custom Object entity.

For more information, see Adobe Marketo Engage Integration Best Practices and Bulk Extract .

Connecting to Amazon Redshift in AWS Glue Studio

> **Note**
>
> You can use AWS Glue for Spark to read from and write to tables in Amazon Redshift databases outside of AWS Glue Studio. To configure Amazon Redshift with AWS Glue jobs programatically, see Redshift connections.

AWS Glue provides built-in support for Amazon Redshift. AWS Glue Studio provides a visual interface to connect to Amazon Redshift, author data integration jobs, and run them on AWS Glue Studio serverless Spark runtime.

Topics

- Creating an Amazon Redshift connection
- Creating a Amazon Redshift source node
- Creating an Amazon Redshift target node
- Advanced options

Creating an Amazon Redshift connection

Permissions needed

Additional permissions are need to use Amazon Redshift clusters and Amazon Redshift serverless environments. For more information on how to add permissions to ETL jobs, see Review IAM permissions needed for ETL jobs.

- redshift:DescribeClusters
- redshift-serverless:ListWorkgroups
- redshift-serverless:ListNamespaces

Overview

When adding an Amazon Redshift connection, you can choose an existing Amazon Redshift connection or create a new connection when adding a **Data source - Redshift** node in AWS Glue Studio.

AWS Glue supports both Amazon Redshift clusters and Amazon Redshift serverless environments. When you create a connection, Amazon Redshift serverless environments display the **serverless** label next to the connection option.

For more information on how to create a Amazon Redshift connection, see [Moving data to and from Amazon Redshift].

Creating a Amazon Redshift source node

Permissions needed

AWS Glue Studio jobs using Amazon Redshift data sources require additional permissions. For more information on how to add permissions to ETL jobs, see [Review IAM permissions needed for ETL jobs].

The following permissions are needed in order to use an Amazon Redshift connection.

- redshift-data:ListSchemas
- redshift-data:ListTables
- redshift-data:DescribeTable
- redshift-data:ExecuteStatement
- redshift-data:DescribeStatement
- redshift-data:GetStatementResult

Adding an Amazon Redshift data source

To add a Data Source – Amazon Redshift node:

1. Choose the Amazon Redshift access type:
   - Direct data connection (recommended) – choose this option if you want to access your Amazon Redshift data directly. This is the recommended option and also the default.
   - Data Catalog tables – choose this option if you have Data Catalog tables that you want to use.
2. If you choose Direct data connection, choose the connection for your Amazon Redshift data source. This assumes that the connection already exists and you can select from existing connections. If you need to create a connection, choose **Create Redshift connection**. For more information, see [Overview of using connectors and connections].

Once you have chosen a connection, you can view the connection properties by clicking **View properties**. Information about the connection are visible, including URL, security groups, subnet, availability zone, description, and created (UTC) and last updated (UTC) timestamps.

3. Choose a Amazon Redshift source option:

    - **Choose a single table** – this is the table that contains the data you want to access from a single Amazon Redshift table.
    - **Enter custom query** – allows you to access a dataset from multiple Amazon Redshift tables based on your custom query.

4. If you chose a single table, choose the Amazon Redshift schema. The list of available schema to choose from is determined by the selected table.

    Or, choose **Enter custom query**. Choose this option to access a custom dataset from multiple Amazon Redshift tables. When you choose this option, enter the Amazon Redshift query.

    When connecting to an Amazon Redshift serverless environment, add the following permission to the custom query:

    ```
                GRANT SELECT ON ALL TABLES IN <schema> TO PUBLIC
    ```

    You can choose **Infer schema** to read the schema based on the query that you entered. You can also choose **Open Redshift query editor** to enter a Amazon Redshift query. For more information, see [Querying a database using the query editor]().

5. In **Performance and security**, choose the Amazon S3 staging directory and IAM role.

    - **Amazon S3 staging directory** – choose the Amazon S3 location for temporarily staging data.
    - **IAM role** – choose the IAM role that can write to the Amazon S3 location you selected.

6. In **Custom Redshift paramters - optional**, enter the parameter and value.

Creating an Amazon Redshift target node

Permissions needed

AWS Glue Studio jobs using Amazon Redshift data target require additional permissions. For more information on how to add permissions to ETL jobs, see Review IAM permissions needed for ETL jobs.

The following permissions are needed in order to use an Amazon Redshift connection.

- redshift-data:ListSchemas
- redshift-data:ListTables

Adding an Amazon Redshift target node

To create a a Amazon Redshift target node:

1. Choose an existing Amazon Redshift table as the target, or enter a new table name.
2. When you use the **Data target - Redshift** target node, you can choose from the following options:

   - **APPEND** – If a table already exists, dump all the new data into the table as an insert. If the table doesn't exist, create it and then insert all new data.

     Additionally, check the box if you want to update (UPSERT) existing records in the target table. The table must exist first, otherwise the operation will fail.
   - **MERGE** – AWS Glue will update or append data to your target table based on the conditions you specify.

     > **Note**
     >
     > To use the merge action in AWS Glue, you must enable Amazon Redshift merge functionality. For instructions on how to enable merge for your Amazon Redshift instance, see MERGE (preview) .

     Choose options:

     - **Choose keys and simple actions** – choose the columns to be used as matching keys between the source data and your target data set.

    Specify the following options when matched:

    - Update record in your target data set with data from source.
    - Delete record in your target data set.

    Specify the following options when not matched:

    - Insert source data as a new row into your target data set.
    - Do nothing.
  - **Enter custom MERGE statement –** You can then choose **Validate Merge statement** to verify that the statement is valid or invalid.
- **TRUNCATE –** If a table already exists, truncate the table data by first clearing the contents of the target table. If truncate is successful, then insert all data. If the table doesn't exist, create the table and insert all data. If truncate is not successful, the operation will fail.
- **DROP –** If a table already exists, delete the table metadata and data. If deletion is successful, then insert all data. If the table doesn't exist, create the table and insert all data. If drop is not successful, the operation will fail.
- **CREATE –** Create a new table with the default name. If table name already exist, create a new table with a name postfix of `job_datetime` to the name for uniqueness. This will insert all the data into the new table. If the table exists, the final table name will have the postfix appended. If the table doesn't exist, a table will be created. In either case, a new table will be created.

Advanced options

See [Using the Amazon Redshift Spark connector on AWS Glue].

Connecting to Asana

Asana is a cloud-based team collaboration solution that helps teams organize, plan, and complete tasks and projects. If you're an Asana user, your account contains data about your workspaces, projects, tasks, teams, and more. You can transfer data from Asana to certain AWS services or other supported applications.

Topics

- AWS Glue support for Asana
- Policies containing the API operations for creating and using connections

- Configuring Asana
- Configuring Asana connections
- Reading from Asana entities
- Asana connection options
- Creating an Asana account
- Limitations

AWS Glue support for Asana

AWS Glue supports Asana as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Asana.

Supported as a target?

No.

Supported Asana API versions

1.0

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the preceding method, alternatively, use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Asana

Before you can use AWS Glue to transfer from Asana, you must meet the following requirements:

Minimum requirements

- You have an Asana account with email and password. For more information about creating an account, see Creating a Asana account.
- You must have AWS Account created with the service access to AWS Glue.
- Ensure you have created one of the following resources in your Asana account:
  - A Developer App that supports `OAuth 2.0` authentication. For more instutions instructions, see OAuth in the Asana Developers documentation. Alternatively, see the section called "Creating an Asana account".
  - A personal access token. For more information, see the Personal access tokenhttps://developers.asana.com/docs/personal-access-token in the Asana Developers documentation.

If you meet these requirements, you're ready to connect AWS Glue to your Adobe Analytics account. For typical connections, you don't need do anything else in Adobe Analytics.

Configuring Asana connections

Asana supports the `AUTHORIZATION_CODE` grant type for `OAuth2`.

This grant type is considered “three-legged” `OAuth` as it relies on redirecting users to the third-party authorization server to authenticate the user. Users may opt to create their own connected app in Asana and provide their own client ID and client secret when creating connections through the AWS Glue console. In this scenario, they will still be redirected to Asana to login and authorize AWS Glue to access their resources.

This grant type results in a refresh token and access token. The access token is short lived, and may be refreshed automatically without user interaction using the refresh token.

For public Asana documentation on creating a connected app for `AUTHORIZATION_CODE OAuth` flow, see Asana APIs .

To configure a Asana connection:

1. In AWS Secrets Manager, create a secret with the following details:
   - For customer managed connected app – Secret should contain the connected app Consumer Secret with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as key.
   - > **Note**
     >
     > You must create a secret for the connection in AWS Glue.
2. In AWS Glue Studio, create a connection under **Data Connections** by following the steps below:
   1. When selecting a **Connection type**, select Asana.
   2. Provide the Asana environment.
   3. Select the IAM role for which AWS Glue can assume and has permissions for following actions:

      JSON

      ```
      {
        "Version": "2012-10-17",
        "Statement": [
          {
            "Effect": "Allow",
      ```

```
        "Action": [
          "secretsmanager:DescribeSecret",
          "secretsmanager:GetSecretValue",
          "secretsmanager:PutSecretValue",
          "ec2:CreateNetworkInterface",
          "ec2:DescribeNetworkInterfaces",
          "ec2:DeleteNetworkInterface"
        ],
        "Resource": "*"
      }
    ]
  }
```

d. Select the `secretName` that you want to use for this connection in AWS Glue to put the tokens.

e. Select the network options if you want to use your network.

3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Asana entities

Prerequisites

An Asana Object you would like to read from. Refer the supported entities table below to check the available entities.

Supported entities for source

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Workspace	No	Yes	No	Yes	No
Tag	No	Yes	No	Yes	No
User	No	Yes	No	Yes	No
Portfolio	No	Yes	No	Yes	No
Team	No	Yes	No	Yes	No
Project	Yes	Yes	No	Yes	No

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Section	No	Yes	No	Yes	No
Task	Yes	No	No	Yes	Yes
Goal	Yes	Yes	No	Yes	No
AuditLogEvent	Yes	Yes	No	Yes	No
Status Update	Yes	Yes	No	Yes	No
Custom Field	No	Yes	No	Yes	No
Project Brief	Yes	No	No	Yes	Yes

Example

```
read_read = glueContext.create_dynamic_frame.from_options(
    connection_type="Asana",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "task/workspace:xxxx",
        "API_VERSION": "1.0",
        "PARTITION_FIELD": "created_at",
        "LOWER_BOUND": "2024-02-05T14:09:30.115Z",
        "UPPER_BOUND": "2024-06-07T13:30:00.134Z",
        "NUM_PARTITIONS": "3"
    }
```

Asana entity and field details

- Workspace
- Tag
- User
- Portfolio

- Team
- Project
- Section
- Task
- Goal
- AuditLogEvent
- Status Update
- Custom Field
- Project Brief

Partitioning queries

Additional spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, `NUM_PARTITIONS` can be provided if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition query.
- `LOWER_BOUND`: an inclusive lower bound value of the chosen partition field.

  For date, we accept the Spark date format used in Spark SQL queries. Example of valid values: `2024-06-07T13:30:00.134Z`.
- `UPPER_BOUND`: an exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: number of partitions.

Entity-wise partitioning field support details are captured in the following table.

Entity Name	Partitioning Field	Data Type
Task	created_at	DateTime
Task	modified_at	DateTime

Example

```
read_read = glueContext.create_dynamic_frame.from_options(
    connection_type="Asana",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "task/workspace:xxxx",
        "API_VERSION": "1.0",
        "PARTITION_FIELD": "created_at",
        "LOWER_BOUND": "2024-02-05T14:09:30.115Z",
        "UPPER_BOUND": "2024-06-07T13:30:00.134Z",
        "NUM_PARTITIONS": "3"
    }
```

Asana connection options

The following are connection options for Asana:

- `ENTITY_NAME`(String) – (Required) Used for Read/Write. The name of your Object in Asana.
- `API_VERSION`(String) – (Required) Used for Read/Write. Asana Rest API version you want to use. For example: 1.0.
- `SELECTED_FIELDS`(List<String>) – Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) – Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) – Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.

Creating an Asana account

1. Sign up for an Asana Account and choose **Sign Up**.
2. After logging in, you will be redirected to the Account Setup page. Complete the following steps:

   - Review the account setup form.
   - Fill in all the relevant details to create your Asana account.
   - Double-check the information for accuracy.
3. Choose **Create Account** or **Submit** (the exact button text may vary) to finalize your account setup.

Creating the App in Asana for `OAuth2.0`

1. Log in to Asana account using your Asana Customer Credentials.
2. Choose your user profile icon in the top-right corner and select **My Settings** from the dropdown menu.
3. Select the **Apps** tab and then select **Manage Developer Apps**.
4. Select **Create new app** and enter the relevant details.
5. Choose **Create Apps**.
6. On the **My Apps** page:
   a. Select **OAuth** and in the **App Credentials** section, make a note of your Client ID and Client Secret.
   b. In the **Redirect URLs** section, add the necessary redirect URL(s).

> **Note**
>
> Enter the Redirect URI using this format: `https://{aws-region-code}.console.aws.amazon.com/gluestudio/oauth`. Example: For the US East (N. Virginia), use: `https://us-east-1.console.aws.amazon.com/gluestudio/oauth`

Creating the App in Asana for PAT Token

1. Log in to Asana account using your Asana Customer Credentials.
2. Choose on your user profile icon in the top-right corner and select **My Profile Settings** from the dropdown menu.
3. Select the **Apps** tab and then select **Service accounts**.

4. Select **Create new app** and enter the relevant details.
5. Choose **Add service account**.
6. The next page displays your token, copy your token and store it securely.

> ⚠ **Important**
>
> This token will only be displayed once. Ensure you copy it and store it securely.

Limitations

The following are limitations for the Asana connector:

- Service Accounts in Enterprise Domains can only access audit log API endpoints. Authentication with a Service Account's personal access token is required to access these endpoints.
- The Goal entity can only be accessed for user accounts with Premium plan or above.
- `Audit Log Event Entity` – In the connector, `start_at` and `end_at` fields are combined into a single field "start_end_at" to support filtering and incremental transfer.
- Partitioning cannot be supported for the `Date` field, even though it supports greater-than-or-equal-to and less-than-or-equal-to operators. Scenario: Created a job with `partitionField` as due_on (datatype: date), `lowerBound` as `2019-09-14`, `upperBound` as `2019-09-16`, and `numPartition` as 2. The filter part of the endpoint URL is created as follows:
  - partition1: due_on.before=2019-09-14&due_on.after=2019-09-14
  - partition2: due_on.before=2019-09-15&due_on.after=2019-09-15 Output:
  - In partition1, we get data with due_date as 2019-09-14 and 2019-09-15
  - In partition2, we get the same data with due_date as 2019-09-15 (which was in partition1) along with other data, causing data duplication.
- Filtering and partitioning cannot be supported on the same field as a bad request error is thrown from the SaaS end.
- The Task entity requires a minimum of 1 field in filter criteria. There is a limitation with Asana where pagination is not identified without sorting the records based on a time-based field. Hence, the Created_at field is used along with pagination to distinguish the next set of records. The Created_at field is marked as mandatory in the filter, with a default value of 2000-01-01T00:00:00Z if not provided. For more information about Pagination, see [Tasks in a workspace](https://developers.asana.com).

Connecting to Azure Cosmos DB in AWS Glue Studio

AWS Glue provides built-in support for Azure Cosmos DB. AWS Glue Studio provides a visual interface to connect to Azure Cosmos DB for NoSQL, author data integration jobs, and run them on the AWS Glue Studio serverless Spark runtime.

Topics

- Creating a Azure Cosmos DB connection
- Creating a Azure Cosmos DB source node
- Creating a Azure Cosmos DB target node
- Advanced options

Creating a Azure Cosmos DB connection

Prerequisites:

- In Azure, you will need to identify or generate an Azure Cosmos DB Key for use by AWS Glue, `cosmosKey`. For more information, see Secure access to data in Azure Cosmos DB in the Azure documentation.

To configure a connection to Azure Cosmos DB:

1. In AWS Secrets Manager, create a secret using your Azure Cosmos DB Key. To create a secret in Secrets Manager, follow the tutorial available in Create an AWS Secrets Manager secret in the AWS Secrets Manager documentation. After creating the secret, keep the Secret name, *`secretName`* for the next step.

    - When selecting **Key/value pairs**, create a pair for the key `spark.cosmos.accountKey` with the value *`cosmosKey`*.

2. In the AWS Glue console, create a connection by following the steps in the section called "Adding an AWS Glue connection". After creating the connection, keep the connection name, *`connectionName`*, for future use in AWS Glue.

    - When selecting a **Connection type**, select Azure Cosmos DB.
    - When selecting an **AWS Secret**, provide *`secretName`*.

Creating a Azure Cosmos DB source node

Prerequisites needed

- A AWS Glue Azure Cosmos DB connection, configured with an AWS Secrets Manager secret, as described in the previous section, the section called "Creating a Azure Cosmos DB connection".
- Appropriate permissions on your job to read the secret used by the connection.
- A Azure Cosmos DB for NoSQL container you would like to read from. You will need identification information for the container.

  An Azure Cosmos for NoSQL container is identified by its database and container. You must provide the database, *`cosmosDBName`*, and container, *`cosmosContainerName`*, names when connecting to the Azure Cosmos for NoSQL API.

Adding a Azure Cosmos DB data source

To add a Data source – Azure Cosmos DB node:

1. Choose the connection for your Azure Cosmos DB data source. Since you have created it, it should be available in the dropdown. If you need to create a connection, choose **Create Azure Cosmos DB connection**. For more information see the previous section, the section called "Creating a Azure Cosmos DB connection".

   Once you have chosen a connection, you can view the connection properties by clicking **View properties**.
2. Choose **Cosmos DB Database Name** – provide the name of the database you want to read from, *`cosmosDBName`*.
3. Choose **Azure Cosmos DB Container** – provide the name of the container you want to read from, *`cosmosContainerName`*.
4. Optionally, choose **Azure Cosmos DB Custom Query** – provide a SQL SELECT query to retrieve specific information from Azure Cosmos DB.
5. In **Custom Azure Cosmos properties**, enter parameters and values as needed.

Creating a Azure Cosmos DB target node

Prerequisites needed

- A AWS Glue Azure Cosmos DB connection, configured with an AWS Secrets Manager secret, as described in the previous section, the section called "Creating a Azure Cosmos DB connection".
- Appropriate permissions on your job to read the secret used by the connection.
- A Azure Cosmos DB table you would like to write to. You will need identification information for the container. **You must create the container before calling the connection method.**

  An Azure Cosmos for NoSQL container is identified by its database and container. You must provide the database, *`cosmosDBName`*, and container, *`cosmosContainerName`*, names when connecting to the Azure Cosmos for NoSQL API.

Adding a Azure Cosmos DB data target

To add a Data target – Azure Cosmos DB node:

1. Choose the connection for your Azure Cosmos DB data source. Since you have created it, it should be available in the dropdown. If you need to create a connection, choose **Create Azure Cosmos DB connection**. For more information see the previous section, the section called "Creating a Azure Cosmos DB connection".

   Once you have chosen a connection, you can view the connection properties by clicking **View properties**.
2. Choose **Cosmos DB Database Name –** provide the name of the database you want to read from, *`cosmosDBName`*.
3. Choose **Azure Cosmos DB Container –** provide the name of the container you want to read from, *`cosmosContainerName`*.
4. In **Custom Azure Cosmos properties**, enter parameters and values as needed.

Advanced options

You can provide advanced options when creating a Azure Cosmos DB node. These options are the same as those available when programming AWS Glue for Spark scripts.

See the section called "Azure Cosmos DB connections".

Connecting to Azure SQL in AWS Glue Studio

AWS Glue provides built-in support for Azure SQL. AWS Glue Studio provides a visual interface to connect to Azure SQL, author data integration jobs, and run them on the AWS Glue Studio serverless Spark runtime.

Topics

- Creating a Azure SQL connection
- Creating a Azure SQL source node
- Creating a Azure SQL target node
- Advanced options

Creating a Azure SQL connection

To connect to Azure SQL from AWS Glue, you will need to create and store your Azure SQL credentials in a AWS Secrets Manager secret, then associate that secret with a Azure SQL AWS Glue connection.

To configure a connection to Azure SQL:

1. In AWS Secrets Manager, create a secret using your Azure SQL credentials. To create a secret in Secrets Manager, follow the tutorial available in Create an AWS Secrets Manager secret in the AWS Secrets Manager documentation. After creating the secret, keep the Secret name, *`secretName`* for the next step.

   - When selecting **Key/value pairs**, create a pair for the key `user` with the value *`azuresqlUsername`*.
   - When selecting **Key/value pairs**, create a pair for the key `password` with the value *`azuresqlPassword`*.

2. In the AWS Glue console, create a connection by following the steps in the section called "Adding an AWS Glue connection". After creating the connection, keep the connection name, *`connectionName`*, for future use in AWS Glue.

   - When selecting a **Connection type**, select Azure SQL.
   - When providing **Azure SQL URL**, provide a JDBC endpoint URL.

The URL must be in the following format: `jdbc:sqlserver://databaseServerName:databasePort;databaseName=azuresqlDBnam`

AWS Glue requires the following URL properties:

- `databaseName` – A default database in Azure SQL to connect to.

For more information about JDBC URLs for Azure SQL Managed Instances, see the Microsoft documentation.

- When selecting an **AWS Secret**, provide *`secretName`*.

Creating a Azure SQL source node

Prerequisites needed

- A AWS Glue Azure SQL connection, configured with an AWS Secrets Manager secret, as described in the previous section, the section called "Creating a Azure SQL connection".
- Appropriate permissions on your job to read the secret used by the connection.
- A Azure SQL table you would like to read from, *`tableName`*.

  An Azure SQL table is identified by its database, schema and table name. You must provide the database name and table name when connecting to Azure SQL. You also must provide the schema if it is not the default, "public". Database is provided through a URL property in *`connectionName`* , schema and table name through the `dbtable`.

Adding a Azure SQL data source

To add a Data source – Azure SQL node:

1. Choose the connection for your Azure SQL data source. Since you have created it, it should be available in the dropdown. If you need to create a connection, choose **Create Azure SQL connection**. For more information see the previous section, the section called "Creating a Azure SQL connection".

   Once you have chosen a connection, you can view the connection properties by clicking **View properties**.

2. Choose a **Azure SQL Source** option:

  - **Choose a single table** – access all data from a single table.
  - **Enter custom query** – access a dataset from multiple tables based on your custom query.
3. If you chose a single table, enter *`tableName`*.

   If you chose **Enter custom query**, enter a TransactSQL SELECT query.
4. In **Custom Azure SQL properties**, enter parameters and values as needed.

Creating a Azure SQL target node

Prerequisites needed

- A AWS Glue Azure SQL connection, configured with an AWS Secrets Manager secret, as described in the previous section, the section called "Creating a Azure SQL connection".
- Appropriate permissions on your job to read the secret used by the connection.
- A Azure SQL table you would like to write to, *`tableName`*.

  An Azure SQL table is identified by its database, schema and table name. You must provide the database name and table name when connecting to Azure SQL. You also must provide the schema if it is not the default, "public". Database is provided through a URL property in *`connectionName`* , schema and table name through the `dbtable`.

Adding a Azure SQL data target

To add a Data target – Azure SQL node:

1. Choose the connection for your Azure SQL data source. Since you have created it, it should be available in the dropdown. If you need to create a connection, choose **Create Azure SQL connection**. For more information see the previous section, the section called "Creating a Azure SQL connection".

   Once you have chosen a connection, you can view the connection properties by clicking **View properties**.
2. Configure **Table name** by providing *`tableName`*.
3. In **Custom Azure SQL properties**, enter parameters and values as needed.

Advanced options

You can provide advanced options when creating a Azure SQL node. These options are the same as those available when programming AWS Glue for Spark scripts.

See the section called "Azure SQL connections".

Connecting to Blackbaud Raiser's Edge NXT

Blackbaud Raiser's Edge NXT is a comprehensive cloud-based fundraising and donor management software solution built specifically for nonprofits and the entire social good community. This connector is built on top of Blackbaud Raiser's Edge NXT's SKY API and provides operations to help manage entities found within the Raisers Edge NXT.

Topics

- AWS Glue support for Blackbaud Raiser's Edge NXT
- Policies containing the API operations for creating and using connections
- Configuring Blackbaud Raiser's Edge NXT
- Configuring Blackbaud Raiser's Edge NXT connections
- Reading from Blackbaud Raiser's Edge NXT entities
- Blackbaud Raiser's Edge NXT connection options
- Blackbaud Raiser's Edge NXT limitations

AWS Glue support for Blackbaud Raiser's Edge NXT

AWS Glue supports Blackbaud Raiser's Edge NXT as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Blackbaud Raiser's Edge NXT.

Supported as a target?

No.

Supported Blackbaud Raiser's Edge NXT API versions

The following Blackbaud Raiser's Edge NXT API versions are supported:

- v1

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Blackbaud Raiser's Edge NXT

Before you can use AWS Glue to transfer data from Blackbaud Raiser's Edge NXT, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- You have a Blackbaud Raiser's Edge NXT account.
- You have generated an Access Token in your Blackbaud Raiser's Edge NXT account with the appropriate read/write scope assigned to access the APIs. For more information, see Authorization.

If you meet these requirements, you’re ready to connect AWS Glue to your Blackbaud Raiser's Edge NXT account.

Configuring Blackbaud Raiser's Edge NXT connections

Blackbaud Raiser's Edge NXT supports the AUTHORIZATION_CODE grant type for OAuth2.

- This grant type is considered "three-legged" OAuth as it relies on redirecting users to a third-party authorization server to authenticate the user. It is used when creating connections via the AWS Glue console. The AWS Glue console will redirect the user to Blackbaud Raiser's Edge NXT where the user must login and allow AWS Glue the requested permissions to access their Blackbaud Raiser's Edge NXT instance.
- Users may opt to create their own connected app in Blackbaud Raiser's Edge NXT and provide their own Client ID, Subscription Key, and Instance URL when creating connections through the AWS Glue console. In this scenario, they will still be redirected to Blackbaud Raiser's Edge NXT to login and authorize AWS Glue to access their resources.
- This grant type results in a refresh token and access token. The access token is short lived, and may be refreshed automatically without user interaction using the refresh token.
- For public Blackbaud Raiser’s Edge NXT documentation on creating a connected app for Authorization Code OAuth flow, see Authorization.

To configure a Blackbaud Raiser's Edge NXT connection:

1. In AWS Secrets Manager, create a secret with the following details:

   a. For the customer managed connected app, the Secret should contain the connected app API key with *`USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET`* as key.
   b. Note: you must create a secret for your connections in AWS Glue.

1. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Data Source**, select Blackbaud Raiser's Edge NXT.
   b. Provide the `INSTANCE_URL` of the Blackbaud Raiser's Edge NXT account you want to connect to.
   c. Provide the user managed client application `clientId`.
   d. Provide the subscription key associated with your account.
   e. Select the AWS IAM role which AWS Glue can assume and has permissions for following actions:

      JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

   f. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.
   g. Select the network options if you want to use your network.
2. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Blackbaud Raiser's Edge NXT entities

Prerequisite

A Blackbaud Raiser's Edge NXT object you would like to read from. You will need the object name.

Supported entities for source:

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Constituent Address	Yes	Yes	No	Yes	Yes
Constituent Education	Yes	Yes	No	Yes	Yes
Constituent Email address	Yes	Yes	No	Yes	Yes
Constituent Phone	Yes	Yes	No	Yes	Yes
Constituent Note	Yes	Yes	No	Yes	Yes
Constituent Relationship	Yes	Yes	No	Yes	Yes
Constituent Online presence	Yes	Yes	No	Yes	Yes
Opportunity	Yes	Yes	No	Yes	Yes
Appeal	Yes	Yes	No	Yes	Yes
Campaign	Yes	Yes	No	Yes	Yes
Fund	Yes	Yes	No	Yes	Yes

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Package	Yes	Yes	No	Yes	Yes
Gift Batch	Yes	Yes	No	Yes	No
Event Participant	Yes	Yes	Yes	Yes	Yes
Constituent Fundraiser Assignment	No	No	No	Yes	No
Gift	Yes	Yes	Yes	Yes	Yes
Membership	Yes	Yes	No	Yes	Yes
Action	Yes	Yes	No	Yes	No
Constituent	Yes	Yes	Yes	Yes	Yes
Constituent Goods	Yes	Yes	No	Yes	Yes
Event	Yes	Yes	Yes	Yes	Yes
Gift custom field	Yes	Yes	No	Yes	Yes

Example:

```
blackbaud_read = glueContext.create_dynamic_frame.from_options(
    connection_type="BLACKBAUD",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "entityName",
        "API_VERSION": "v1",
        "SUBSCRIPTION_KEY": <Subscription key associated with one's developer account>
    }
```

Blackbaud Raiser's Edge NXT entity and field details

For more information about the entities and field details see:

- Action
- Constituent
- Constituent Address
- Constituent Membership
- Constituent Fundraiser Assignment
- Constituent Education
- Constituent Email Address
- Constituent Phone
- Constituent Note
- Constituent Online Presence
- Constituent Relationship
- Event
- Event Participant
- Appeal
- Campaign
- Fund
- Package
- Gift
- Gift Custom Field
- Gift Batch
- Opportunity
- Constituent Codes

Note

Struct and List data types are converted to String data type, and DateTime data type is converted to Timestamp in the response of the connectors.

Partitioning queries

Field-based partitioning:

Blackbaud Raiser's Edge NXT doesn’t support field based or record based partitioning.

Record-based partitioning:

You can provide the additional Spark option `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With this parameter, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

In record based partitioning, the total number of records present is queried from Blackbaud Raiser’s Edge NXT API, and it is divided by `NUM_PARTITIONS` number provided. The resulting number of records are then concurrently fetched by each sub-query.

- `NUM_PARTITIONS`: the number of partitions.

Example:

```
blackbaud_read = glueContext.create_dynamic_frame.from_options(
    connection_type="BLACKBAUD",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "entityName",
        "API_VERSION": "v1",
        "NUM_PARTITIONS": "2",
        "SUBSCRIPTION_KEY": <Subscription key associated with one's developer account>
    }
```

Blackbaud Raiser's Edge NXT connection options

The following are connection options for Blackbaud Raiser's Edge NXT:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your object in Blackbaud Raiser's Edge NXT.
- `API_VERSION`(String) - (Required) Used for Read. Blackbaud Raiser's Edge NXT Rest API version you want to use.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.

- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read. Example value: 10.
- `SUBSCRIPTION_KEY`(String) - (Required) Default: empty. Used for Read. Subscription key associated with one's developer account.

Blackbaud Raiser's Edge NXT limitations

The following are limitations or notes for Blackbaud Raiser's Edge NXT:

- The SaaS only supports the `EQUAL_TO` operator, which returns results created or modified on or after the specified date. Additionally, the "id" field is a String data type. There is also no identification of non-nullable fields. Therefore, field-based partitioning is not supported.
- Incremental pull is only supported by the `Event` entity with daily, monthly and weekly frequencies.
- The Constituent Fundraiser Assignment entity returns a maximum of 20 records.
- Record-based partitioning:
  - Not supported by the `Action`, `Constituent Fundraiser Assignment` or `Gift Batch` entities.
  - Record-based partitioning with the filter predicate is only supported by the `Event` and `Event Participant` entities. If a filter predicate is used with any other record-based supported entities, an exception will be thrown.
- In the `Gift Custom Field` entity, the field 'value' must be used in conjunction with the field 'category', which otherwise leads to an unfiltered response. Thus, to compel the user to plug in the 'category' field while filtering with the 'value' field, an exception will be thrown if the aforementioned requirement has not been followed.
- The `date_added` and `last_modified` fields for all applicable entities do not support any comparative operators. They only support the equal to operator. Also, there is no field that can be paired with the aforementioned fields to provide a range of records. Hence, these fields are only queryable and cannot support incremental transfer.
- The `added_by` field in the `Gift Batch` entity will not be considered as filterable as it might not emit the correct results.

- There is a latency of approximately 30 minutes for records to be retrieved via the `/GET Gift List` endpoint upon insertion of data in the `Gift` entity.
- Support for incremental transfer has been dropped for the Gift entity due to limitations from the data source's end.
- There exists a 10 minute latency for the status field in the Opportunity entity.
- The `Fundraiser Assignment` entity has `Constituent` as the dependent entity. The connector loads at most 5,000 IDs to choose from, to avoid the response size exceeding the maximum allowed payload size.

Connecting to CircleCI

CircleCI is a continuous integration and continuous delivery platform. Your CircleCI account contains data about your projects, pipelines, workflows, and more. If you're a CircleCI user, you can connect AWS Glue to your CircleCI account. Then, you can use CircleCI as a data source in your ETL jobs. Run these jobs to transfer data between CircleCI and AWS services or other supported applications.

Topics

- AWS Glue support for CircleCI
- Policies containing the API operations for creating and using connections
- Configuring CircleCI
- Configuring CircleCI connections
- Reading from CircleCI entities
- CircleCI connection options
- CircleCI limitations

AWS Glue support for CircleCI

AWS Glue supports CircleCI as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from CircleCI.

Supported as a target?

No.

Supported CircleCI API versions

The following CircleCI API versions are supported:

- v2

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.

- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring CircleCI

Before you can use AWS Glue to transfer data from CircleCI, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- You have an account with CircleCI that contains the data that you want to transfer.
- In the user settings for your account, you've created a personal API token. For more information, see Creating a personal API token.
- You provide the personal API token to AWS Glue while creating the connection.

If you meet these requirements, you’re ready to connect AWS Glue to your CircleCI account.

Configuring CircleCI connections

CircleCI supports custom authentication.

To configure a CircleCI connection:

1. In AWS Secrets Manager, create a secret with the following details:
   a. For the customer managed connected app, the Secret should contain the connected app API key with `Circle-Token` as key.
   b. Note: you must create a secret for your connections in AWS Glue.

1. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Data Source**, select CircleCI.
   b. Select the AWS IAM role which AWS Glue can assume and has permissions for following actions:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

c. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.

d. Select the network options if you want to use your network.

2. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from CircleCI entities

Prerequisite

A CircleCI object you would like to read from. You will need the object name.

Supported entities for source:

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Context	Yes	No	No	Yes	No

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Organization Summary Metric	Yes	No	No	Yes	No
Pipeline	No	No	No	Yes	No
Pipeline Workflow	Yes	No	No	Yes	No
Project Branch	Yes	No	No	Yes	No
Project Flaky Test	No	No	No	Yes	No
Project Summary Metric	Yes	No	No	Yes	No
Schedule	No	No	No	Yes	No
Workflow Job Timeseries	Yes	No	No	Yes	No
Workflow Metric And Trend	Yes	No	No	Yes	No
Workflow Recent Run	Yes	No	No	Yes	No
Workflow Summary Metric	Yes	No	No	Yes	No

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Workflow Test Metric	Yes	No	No	Yes	No

Example:

```
circleci_read = glueContext.create_dynamic_frame.from_options(
    connection_type="circleci",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "context/e7ea2945-dccb-4205-b673-8391fe1b3a4c",
        "API_VERSION": "v2"
    }
```

CircleCI entity and field details

For more information about the entities and field details see:

- Contexts
- Project Summary Metrics
- Workflow Job Timeseries
- Organization Summary Metrics
- Project Branches
- Project Flaky Tests
- Workflow Recent Runs
- Workflow Summary Metrics
- Workflow Metrics and Trends
- Workflow Test Metrics
- Pipelines
- Pipeline Workflows
- Schedules

Entities with static metadata:

Entity	Field	Data type	Supported operators
Context	Created At	String	
	ID	String	
	Name	String	
	Owner Type	String	EQUAL_TO
Organization Summary Metric	All Projects	List	
	Org Data	Struct	
	Org Project Data	List	
	Project Names	String	EQUAL_TO
	Reporting Window	String	EQUAL_TO
Pipeline	Branch	String	EQUAL_TO
	Created At	String	
	Errors	List	
	ID	String	
	Number	Integer	
	Project Slug	String	
	State	String	
	Trigger	Struct	
	Trigger Parameters	Struct	
	Updated At	String	
	VCS	Struct	

Entity	Field	Data type	Supported operators
Pipeline Workflow	Canceled By	String	
	Created At	String	
	Errorer By	String	
	ID	String	
	Name	String	
	Pipeline ID	String	
	Pipeline Number	Integer	
	Project Slug	String	
	Started By	String	
	Status	String	
	Stopped At	String	
	Tag	String	
Project Branch	Branches	List	
	Org ID	String	
	Project ID	String	
	Workflow Name	String	EQUAL_TO
Project Flaky Test	Classname	String	
	File	String	
	Job Name	String	
	Job Number	Integer	

Entity	Field	Data type	Supported operators
	Pipeline Number	Integer	
	Source	String	
	Test Name	String	
	Time Wasted	Integer	
	Times Flaked	Integer	
	Workflow Created At	String	
	Workflow ID	String	
	Workflow Name	String	
Project Summary Metric	All Branches	List	
	All Workflows	List	
	Branches	String	EQUAL_TO
	Organization ID	String	
	Project Data	Struct	
	Project ID	String	
	Project Workflow Branch Data	List	
	Project Workflow Data	List	
	Reporting Window	String	EQUAL_TO
	Workflow Names	String	EQUAL_TO
Schedule	Actor	Struct	

Entity	Field	Data type	Supported operators
	Created At	String	
	Description	String	
	ID	String	
	Name	String	
	Parameters	Struct	
	Project Slug	String	
	Timetable	Struct	
	Updated At	String	
Workflow Job Timeseries	Branch	String	EQUAL_TO
	Granularity	String	EQUAL_TO
	Max Ended At	String	
	Metrics	Struct	
	Min Started At	String	
	Name	String	
	Start End Date	DateTime	EQUAL_TO, BETWEEN
	Timestamp	String	
Workflow Metric and Trend	All Branches	Boolean	EQUAL_TO
	Branches	String	EQUAL_TO
	Metrics	Struct	
	Trends	Struct	

Entity	Field	Data type	Supported operators
	Workflow Names	List	
Workflow Recent Run	All Brances	Boolean	EQUAL_TO
	Branch	String	EQUAL_TO
	Created At	String	
	Credits Used	Integer	
	Duration	Integer	
	ID	String	
	Is Approval	Boolean	
	Start End Date	DateTime	EQUAL_TO, BETWEEN
	Status	String	
	Stopped At	String	
Workflow Summary Metric	All Branches	Boolean	EQUAL_TO
	Branch	String	EQUAL_TO
	Metrics	Struct	
	Name	String	
	Project ID	String	
	Reporting Window	String	EQUAL_TO
	Window End	String	
	Window Start	String	
Workflow Test Metric	Average Test Count	Integer	

Entity	Field	Data type	Supported operators
	Branch	String	EQUAL_TO
	Most Failed Tests	List	
	Most Failed Tests Extra	Integer	
	Slowest Tests	List	
	Slowest Tests Extra	Integer	
	Test Runs	List	
	Total Test Runs	Integer	

> **Note**
>
> Struct and List data types are converted to String data type in the response of the connector.

Partitioning queries

CircleCI doesn’t support field-based or record-based partitioning.

CircleCI connection options

The following are connection options for CircleCI:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your object in CircleCI.
- `API_VERSION`(String) - (Required) Used for Read. CircleCI Rest API version you want to use.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.

CircleCI limitations

The following are limitations or notes for CircleCI:

- CircleCI does not support either field based or record based partitioning.
- Filter fields containing '-' (hyphen) will work only if they are wrapped within backticks. For example: `` `workflow-name` = "abc" ``
- The GitLab VCS type cannot be supported as there is no programmatic way to retrieve the 'Project ID' required for the GitLab VCS entity path.

Connecting to Datadog

Datadog is a monitoring and analytics platform for cloud-scale applications, including infrastructure, applications, services, and tools.

Topics

- AWS Glue support for Datadog
- Policies containing the API operations for creating and using connections
- Configuring Datadog
- Configuring Datadog connections
- Reading from Datadog entities
- Datadog connection options
- Creating a Datadog account
- Limitations

AWS Glue support for Datadog

AWS Glue supports Datadog as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Datadog.

Supported as a target?

No.

Supported Datadog API versions

- v1
- v2

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the preceding method, alternatively, use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management console. If you follow the naming convention

for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Datadog

Before you can use AWS Glue to transfer from Datadog, you must meet the following requirements:

Minimum requirements

- You have a Datadog account with DD-API-KEY and DD-APPLICATION-KEY. For more information about creating an account, see Creating a Datadog account.
- Your Datadog account has API access with valid license.

Datadog supports the following six URLs. All Datadog API clients are configured by default to consume Datadog US1 site APIs. If you are on the Datadog EU site, you must select https://api.datadoghq.eu URL with the `DD-API-KEY` and `DD-APPLICATION-KEY` of the Datadog EU site to access the APIs. Similarly, for other sites, you should select the respective URLs with the `DD-API-KEY and DD-APPLICATION-KEY` of the respective site.

- US1 API URL — https://api.datadoghq.comhttps://api.datadoghq.com
- EU API URL — https://api.datadoghq.eu
- US3 API URL — https://api.us3.datadoghq.com
- US5 API URL — https://api.us5.datadoghq.com
- S1-FED API URL — https://api.ddog-gov.com
- Japan API URL — https://api.ap1.datadoghq.com

If you meet these requirements, you’re ready to connect AWS Glue to your Datadog account.

Configuring Datadog connections

Datadog supports custom authentication. Following are the steps to configure Datadog connection:

To configure a Datadog connection:

1. In AWS Secrets Manager, create a secret with the following details:

   For customer managed connected app – Secret should contain the connected app Consumer Secret with **`API_KEY`** and **`APPLICATION_KEY`** as keys.

   > **Note**
   >
   > It is a must to create a secret per connection in AWS Glue.

2. In AWS Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Connection type**, select Datadog.
   b. Provide the `Instance_Url` of the Datadog you want to connect to.
   c. Select the IAM role for which AWS Glue can assume and has permissions for following actions:

      JSON

      ```
      {
        "Version": "2012-10-17",
        "Statement": [
          {
            "Effect": "Allow",
            "Action": [
              "secretsmanager:DescribeSecret",
              "secretsmanager:GetSecretValue",
              "secretsmanager:PutSecretValue",
              "ec2:CreateNetworkInterface",
              "ec2:DescribeNetworkInterfaces",
              "ec2:DeleteNetworkInterface"
            ],
            "Resource": "*"
          }
        ]
      }
      ```

   d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.
   e. Select the network options if you want to use your network.

3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Datadog entities

Prerequisites

A Datadog Object you would like to read from. Refer the supported entities table below to check the available entities.

Supported entities

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Metrics Timeseries	Yes	No	No	Yes	No
Log Queries	Yes	Yes	Yes	Yes	No

Example

```
Datadog_read = glueContext.create_dynamic_frame.from_options(
    connection_type="datadog",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "log-queries",
        "API_VERSION": "v2",
        "INSTANCE_URL": "https://api.datadoghq.com",
        "FILTER_PREDICATE": "from = `2023-10-03T09:00:26Z`"
    }
```

Datadog entity and field details

Entity	Field	Data Type	Supported Operators
Metrics Timeseries	error	String	NA
	aggr	String	NA
	attributes	Struct	NA

Entity	Field	Data Type	Supported Operators
	display_name	String	NA
	end	DateTime	NA
	expression	String	NA
	interval	Integer	NA
	length	Integer	NA
	metric	String	NA
	pointlist	List	NA
	query_index	Integer	NA
	scope	String	NA
	start	DateTime	NA
	tag_set	List	NA
	unit	Struct	NA
	from_to_date	DateTime	BETWEEN
	query	String	EQUAL_TO
	status	String	NA
	type	String	NA
	host	String	NA
Log Queries	id	String	NA
	attributes	Struct	NA
	timestamp	DateTime	NA

Entity	Field	Data Type	Supported Operators
	type	String	NA
	from	DateTime	BETWEEN,EQUAL_TO
	indexes	List	EQUAL_TO
	storage_tier	String	EQUAL_TO
	query	String	EQUAL_TO

Datadog connection options

The following are connection options for Datadog:

- `ENTITY_NAME`(String) – (Required) Used for Read/Write. The name of your Object in Datadog.
- `API_VERSION`(String) – (Required) Used for Read/Write. Datadog Rest API version you want to use. `v1` version supports `metrics-timeseries` entity whereas, `v2` version supports `log-queries` entity.
- `INSTANCE_URL`(String) – (Required) Used for Read. Datadog instance URL. Datadog instance URL varies per region.
- `SELECTED_FIELDS`(List<String>) – Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) – Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) – Default: empty. Used for Read. Full Spark SQL query.

Creating a Datadog account

1. Go to https://www.datadoghq.com/.
2. Choose **GET STARTED FREE**.
3. Enter the required information and sign up.
4. Install the **Datadog Agent Installer** as suggested.

5. Ensure that the account gets registered with a valid organization (from the list available) that has an active Datadog subscription.
6. After logging in to your Datadog account, hover over your username in the top-right corner to view the **Keys** details:
    a. To get your API key, choose **API Keys**.
    b. To get your application key, choose **Application Keys**.

Limitations

The following are limitations for the Datadog connector:

- Datadog doesn’t support either field based or record based partitioning.
- `from` is mandatory filter parameter for `Log Queries` entity.
- `from_to_date` and `query` are mandatory filter parameters for `Metrics Timeseries` entity.

Connecting to Docusign Monitor

Docusign Monitor helps organizations protect their agreements with round-the-clock activity tracking. The Monitor API delivers this activity tracking information directly to existing security stacks or data visualization tools—enabling teams to detect unauthorized activity, investigate incidents, and quickly respond to verified threats. It also provides the flexibility, security teams need to customize dashboards and alerts to meet specific business needs.

Topics

- AWS Glue support for Docusign Monitor
- Policies containing the API operations for creating and using connections
- Configuring Docusign Monitor
- Configuring Docusign Monitor connections
- Reading from Docusign Monitor entities
- Docusign Monitor connection options
- Docusign Monitor limitations

AWS Glue support for Docusign Monitor

AWS Glue supports Docusign Monitor as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Docusign Monitor.

Supported as a target?

No.

Supported Docusign Monitor API versions

The following Docusign Monitor API versions are supported:

- v2.0

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```json
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Docusign Monitor

Before you can use AWS Glue to transfer data from Docusign Monitor to supported destinations, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- You have an Docusign account where you use the Docusign Software product in Docusign Monitor.
- In the developer console for your Docusign account, you've created an OAuth 2.0 integration app for AWS Glue.

  This app provides the client credentials that AWS Glue uses to access your data securely when it makes authenticated calls to your account. For more information, see OAuth 2.0 in the Docusign Monitor documentation.

If you meet these requirements, you’re ready to connect AWS Glue to your Docusign Monitor account.

Configuring Docusign Monitor connections

Docusign Monitor supports the AUTHORIZATION_CODE grant type.

- This grant type is considered three-legged OAuth as it relies on redirecting users to the third-party authorization server to authenticate the user. It is used when creating connections via the AWS Glue console.
- Users may opt to create their own connected app in Docusign Monitor and provide their own client ID and client secret when creating connections through the AWS Glue console. In this scenario, they will still be redirected to Docusign Monitor to login and authorize AWS Glue to access their resources.
- This grant type results in a refresh token and an access token. The access token is short lived, and may be refreshed automatically without user interaction using the refresh token.
- For public Docusign Monitor documentation on creating a connected app for the Authorization Code OAuth flow, see OAuth for Docusign Connect.

To configure a Docusign Monitor connection:

1. In AWS Secrets Manager, create a secret with the following details:
   a. For the customer managed connected app, the Secret should contain the connected app API key with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as key.
   b. Note: you must create a secret for your connections in AWS Glue.

1. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. Under **Connections**, choose **Create connection**.
   b. When selecting a **Data Source**, select Docusign Monitor.
   c. Select the AWS IAM role which AWS Glue can assume and has permissions for following actions:

      JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
          "ec2:CreateNetworkInterface",
          "ec2:DescribeNetworkInterfaces",
          "ec2:DeleteNetworkInterface"
        ],
        "Resource": "*"
      }
    ]
  }
```

   d. Provide the **User Managed Client Application ClientId** of the Docusign Monitor app.
   e. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.
   f. Select the network options if you want to use your network.
2. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Docusign Monitor entities

Prerequisite

A Docusign Monitor object you would like to read from.

Supported entities for source:

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Monitoring Data	Yes	Yes	No	Yes	No

Example:

```
docusignmonitor_read = glueContext.create_dynamic_frame.from_options(
    connection_type="docusign_monitor",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "monitoring-data",
        "API_VERSION": "v2.0"
    }
```

Docusign Monitor entity and field details

Entities with static metadata:

Entity	Field	Data type	Supported operators
Monitoring Data	timestamp	DateTime	N/A
	eventId	String	N/A
	application	String	N/A
	environment	String	N/A
	site	String	N/A
	traceToken	String	N/A
	organizationId	String	N/A
	accountId	String	N/A
	userId	String	N/A
	object	String	N/A
	action	String	N/A
	property	String	N/A
	field	String	N/A
	result	String	N/A
	IntegratorKey	String	N/A
	customerVisible	String	N/A
	version	String	N/A
	userAgent	String	N/A

Entity	Field	Data type	Supported operators
	userAgentClientInfo	Struct	N/A
	ipAddress	String	N/A
	ipAddressLocation	Struct	N/A
	data	String	N/A
	source	String	N/A
	latitude	Double	N/A
	longitude	Double	N/A
	city	String	N/A
	state	String	N/A
	country	String	N/A
	usUserMemberOfDomain	Boolean	N/A
	affectedUserIsMemberOfDomain	Boolean	N/A
	proxyStatus	String	N/A
	proxyType	String	N/A
	proxyLevel	String	N/A
	referencedUserId	String	N/A
	device	String	N/A
	browser	String	N/A
	cursor	DateTime	EQUAL_TO

Partitioning queries

Docusign Monitor doesn't support either field-based or record-based partitioning.

Docusign Monitor connection options

The following are connection options for Docusign Monitor:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your object in Docusign Monitor.
- `API_VERSION`(String) - (Required) Used for Read. Docusign Monitor Rest API version you want to use.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.

Docusign Monitor limitations

The following are limitations or notes for Docusign Monitor:

- When a filter is applied using the `cursor` field, the API retrieves records for the next seven days starting from the specified date.
- If no filter is provided, the API retrieves records for the previous seven days from the current date of the API request.
- Docusign Monitor does not support either field-based or record-based partitioning.
- Docusign Monitor does not support the Order By feature.

Connecting to Domo

Domo is a cloud-based dash-boarding tool. With Domo's enterprise application platform, the foundation you need to extend Domo is in place, so you can build your custom solutions faster.

Topics

- AWS Glue support for Domo
- Policies containing the API operations for creating and using connections

- [Configuring Domo]
- [Configuring Domo connections]
- [Reading from Domo entities]
- [Domo connection options]
- [Domo limitations]

AWS Glue support for Domo

AWS Glue supports Domo as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Domo.

Supported as a target?

No.

Supported Domo API versions

The following Domo API versions are supported:

- v1

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
          "glue:ListEntities",
          "glue:DescribeEntity"
        ],
        "Resource": "*"
      }
    ]
  }
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Domo

Before you can use AWS Glue to transfer data from Domo to supported destinations, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- You have a Domo account enabled for API access.
- You have an app under your Domo developer account that provides the client credentials that AWS Glue uses to access your data securely when it makes authenticated calls to your account. For more information, see Creating a Domo developer app.

If you meet these requirements, you’re ready to connect AWS Glue to your Domo account.

Creating a Domo developer app

To get the Client ID and Client Secret you create a developer account.

1. Go to the [Domo developer login page.](#)
2. Choose **Login**.
3. Provide the domain name and click **Continue**.
4. Hover on **My Account** and choose **New Client**.
5. Provide the Name and Description and select the scope ("data") and choose **Create**.
6. Retrieve the generated **Client Id** and **Client Secret** from the new client created.

Configuring Domo connections

Domo supports the CLIENT_CREDENTIALS grant type for OAuth2.

- This grant type is considered two-legged OAuth as only the client application authenticates itself to the server, with no involvement to the user.
- Users may opt to create their own connected app in Domo and provide their own client ID and client secret when creating connections through the AWS Glue console.
- For public Domo documentation on creating a connected app for the Authorization Code OAuth flow, see [OAuth Authentication](#).

To configure a Domo connection:

1. In AWS Secrets Manager, create a secret with the following details:
   a. For the customer managed connected app, the Secret should contain the connected app access token, `client_id`, and `client_secret`.
   b. Note: you must create a secret for your connections in AWS Glue.

1. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. Under **Connections**, choose **Create connection**.
   b. When selecting a **Data Source**, select Domo.
   c. Select the AWS IAM role which AWS Glue can assume and has permissions for following actions:

      JSON

      ```
      {
      ```

```
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

   d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.
   e. Select the network options if you want to use your network.

2. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Domo entities

Prerequisite

A Domo object you would like to read from. You will need the object name such as Data Set or Data Permission Policies. The following table shows the supported entities.

Supported entities for source:

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Data Set	Yes	Yes	Yes	Yes	Yes
Data Permission Policies	No	No	No	Yes	No

Example:

```
Domo_read = glueContext.create_dynamic_frame.from_options(
    connection_type="domo",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "dataset",
        "API_VERSION": "v1"
    }
```

Domo entity and field details

Entities with static metadata:

Entity	Field	Data type	Supported operators
Data Permission Policies	id	Long	N/A
	type	String	N/A
	name	String	N/A
	filters	List	N/A
	users	List	N/A
	virtualUsers	List	N/A
	groups	List	N/A

For the following entity, Domo provides endpoints to fetch metadata dynamically, so that operator support is captured at the datatype level for the entity.

Entity	Data type	Supported operators
Data Set	Integer	=, !=, <, >, >=, <=
	Long	=, !=, <, >, >=, <=
	String	=, !=, CONTAINS

Entity	Data type	Supported operators
	Date	=, >, >=, <, <=, BETWEEN
	DateTime	=, >, >=, <, <=, BETWEEN
	Boolean	=, !=
	Double	=, !=, <, >, >=, <=
	List	N/A
	Struct	N/A

Partitioning queries

Field-based partitioning

You can provide the additional Spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, and `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition the query.
- `LOWER_BOUND`: an **inclusive** lower bound value of the chosen partition field.

  For the DateTime field, we accept the value in ISO format.

  Example of valid value:

  ```
  "2023-01-15T11:18:39.205Z"
  ```

  For the Date field, we accept the value in ISO format.

  Example of valid value:

  ```
  "2023-01-15"
  ```
- `UPPER_BOUND`: an **exclusive** upper bound value of the chosen partition field.

Example of valid value:

```
"2023-02-15T11:18:39.205Z"
```

- `NUM_PARTITIONS`: the number of partitions.

Entity-wise partitioning field support details are captured in the following table:

Entity name	Partitioning fields	Data type
Dataset	Any Date/Time based field [dynamic metadata]	DateTime
	Any Date based field [dynamic metadata]	Date

Example:

```
Domo_read = glueContext.create_dynamic_frame.from_options(
    connection_type="domo",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "dataset",
        "API_VERSION": "v1",
        "PARTITION_FIELD": "permissionTime"
        "LOWER_BOUND": "2023-01-15T11:18:39.205Z"
        "UPPER_BOUND": "2023-02-15T11:18:39.205Z"
        "NUM_PARTITIONS": "2"
    }
```

Record-based partitioning

You can provide the additional Spark option `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With this parameter, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

In record based partitioning, the total number of records present is queried from Domo, and it is divided by the `NUM_PARTITIONS` number provided. The resulting number of records are then concurrently fetched by each sub-query.

Example:

```
Domo_read = glueContext.create_dynamic_frame.from_options(
    connection_type="domo",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "dataset",
        "API_VERSION": "v1",
        "NUM_PARTITIONS": "2"
    }
```

Domo connection options

The following are connection options for Domo:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your object in Domo.
- `API_VERSION`(String) - (Required) Used for Read. Domo Rest API version you want to use.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for read. Number of partitions for read.

Domo limitations

The following are limitations or notes for Domo:

- Due to an SDK limitation, filtration does not work as expected for the queryable fields that starts with '_' (for example: _BATCH_ID).

- Due to an API limitation, filtration works on the date prior to the date you provide. This also affects incremental pull. To overcome this limitation, select a date according to your time zone against UTC, for getting data for the required date.

Connecting to Dynatrace

Dynatrace is a platform that offers analytics and automation for comprehensive observability and security. It specializes in monitoring and optimizing application performance, infrastructure, and user experience.

Topics

- AWS Glue support for Dynatrace
- Policies containing the API operations for creating and using connections
- Configuring Dynatrace
- Configuring Dynatrace connections
- Reading from Dynatrace entities
- Dynatrace connection options
- Dynatrace limitations

AWS Glue support for Dynatrace

AWS Glue supports Dynatrace as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Dynatrace.

Supported as a target?

No.

Supported Dynatrace API versions

The following Dynatrace API versions are supported:

- v2

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Dynatrace

Before you can use AWS Glue to transfer data from Dynatrace, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- You have a Dynatrace account.
- You have generated an Access Token in your Dynatrace account with the appropriate read/write scope assigned to access the APIs. For more information, see Generate a token.

If you meet these requirements, you're ready to connect AWS Glue to your Dynatrace account.

Configuring Dynatrace connections

Dynatrace supports custom authentication.

To configure a Dynatrace connection:

1. In AWS Secrets Manager, create a secret with the following details:
    a. For the customer managed connected app, the Secret should contain the connected app API key with *`apiToken`* as key.
    b. Note: you must create a secret for your connections in AWS Glue.

1. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
    a. When selecting a **Data Source**, select Dynatrace.
    b. Provide the `INSTANCE_URL` of the Dynatrace account you want to connect to.
    c. Select the AWS IAM role which AWS Glue can assume and has permissions for following actions:

    JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
                "ec2:CreateNetworkInterface",
                "ec2:DescribeNetworkInterfaces",
                "ec2:DeleteNetworkInterface"
            ],
            "Resource": "*"
        }
    ]
}
```

   d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.
   e. Select the network options if you want to use your network.

2. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Dynatrace entities

Prerequisite

A Dynatrace object you would like to read from. You will need the object name such as "problem".

Supported entities for source:

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Problem	Yes	Yes	Yes	Yes	No

Example:

```
Dynatrace_read = glueContext.create_dynamic_frame.from_options(
    connection_type="Dynatrace",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "problem",
        "API_VERSION": "v2",
        "INSTANCE_URL": "https://[instanceName].live.dynatrace.com"
    }
```

Dynatrace entity and field details:

Dynatrace provides endpoints to fetch metadata dynamically for supported entities. Accordingly, operator support is captured at the datatype level.

Entity	Field	Data type	Supported operators
Problem	affectedEntities	List	EQUAL_TO
	displayId	String	EQUAL_TO
	endTime	DateTime	
	entityTags	List	
	evidenceDetails	Struct	
	impactAnalysis	Struct	
	impactLevel	String	EQUAL_TO
	impactedEntities	List	EQUAL_TO
	linkedProblemInfo	Struct	
	managementZones	List	EQUAL_TO
	problemFilters	List	
	recentComments	Struct	
	rootCauseEntity	Struct	EQUAL_TO
	problemId	String	EQUAL_TO
	severityLevel	String	EQUAL_TO
	startTime	DateTime	BETWEEN
	status	String	EQUAL_TO
	title	String	

Entity	Field	Data type	Supported operators
	from	DateTime	EQUAL_TO, BETWEEN
	problemFilterIds	String	EQUAL_TO
	problemFilterNames	String	EQUAL_TO
	managementZoneIds	String	EQUAL_TO
	text	String	EQUAL_TO
	underMaintenance	Boolean	EQUAL_TO
	message	String	

Partitioning queries

Dynatrace doesn’t support field based or record based partitioning.

Dynatrace connection options

The following are connection options for Dynatrace:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your object in Dynatrace.
- `API_VERSION`(String) - (Required) Used for Read. Dynatrace Rest API version you want to use.
- `INSTANCE_URL`(String) - Used for Read. A valid Dynatrace instance URL.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.

Dynatrace limitations

The following are limitations or notes for Dynatrace:

- Dynatrace doesn’t support either field based or record based partitioning.
- For the Select All feature, if you provide the "field" in the filter then it will not allow records to be more then 10 per page.
- The maximum page size supported is 500. If you select any of the [`evidenceDetails, impactAnalysis, recentComments`] fields while creating the flow then records per page will be defaulted to 10.

Connecting to Facebook Ads

Facebook Ads is a powerful digital advertising platform used by businesses of all sizes to reach their target audience and achieve various marketing objectives. The platform allows advertisers to create tailored ads that can be displayed across Facebook's family of apps and services, including Facebook and Messenger. With its advanced targeting capabilities, Facebook Ads enables businesses to reach specific demographics, interests, behaviors, and locations.

Topics

- AWS Glue support for Facebook Ads
- Policies containing the API operations for creating and using connections
- Configuring Facebook Ads
- Configuring Facebook Ads connections
- Reading from Facebook Ads entities
- Facebook Ads connection options
- Limitations and notes for Facebook Ads connector

AWS Glue support for Facebook Ads

AWS Glue supports Facebook Ads as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Facebook Ads.

Supported as a target?

No.

Supported Facebook Ads API versions

The following Facebook Ads API versions are supported:

- v17.0
- v18.0
- v19.0
- v20.0

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.

- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Facebook Ads

Before you can use AWS Glue to transfer data from Facebook Ads, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- Facebook Standard accounts are accessed directly through Facebook.
- User authentication is needed to generate the access token.
- The Facebook Ads SDK connector will be implementing the *User Access Token OAuth* flow.
- We are using OAuth2.0 to authenticate our API requests to Facebook Ads. This web-based authentication falls under the Multi-Factor Authentication (MFA) architecture, which is a superset of 2FA.
- The user needs to grant permissions to access the end points. For accessing the user's data, endpoint authorization is handled through permissions and features.

Getting OAuth 2.0 credentials

To obtain API credentials so that you can make authenticated calls to your instance, see REST API in the Facebook Ads Developer Guide.

Configuring Facebook Ads connections

Facebook Ads supports the AUTHORIZATION_CODE grant type for OAuth2.

- This grant type is considered three-legged OAuth as it relies on redirecting users to the third-party authorization server to authenticate the user. It is used when creating connections via the AWS Glue console.
- Users may still opt to create their own connected app in Facebook Ads and provide their own client ID and client secret when creating connections through the AWS Glue console. In this

scenario, they will still be redirected to Facebook Ads to login and authorize AWS Glue to access their resources.

- This grant type results in an access token. An expiring system user token is valid for 60 days from a generated or refreshed date. To create continuity, the developer should refresh the access token within 60 days. Failing to do so results in forfeiting the access token and requires the developer obtain a new one to regain API access. See Refresh Access Token.
- For public Facebook Ads documentation on creating a connected app for Authorization Code OAuth flow, see Using OAuth 2.0 to Access Google APIs in the Google for Developers guide.

To configure a Facebook Ads connection:

1. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Connection type**, select Facebook Ads.
   b. Provide the `INSTANCE_URL` of the Facebook Ads instance you want to connect to.
   c. Select the AWS IAM role which AWS Glue can assume and has permissions for the following actions:

      JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

   d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.
   e. Select the network options if you want to use your network.

2. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Facebook Ads entities

Prerequisite

A Facebook Ads object you would like to read from. You will need the object name. The following tables shows the supported entities.

Supported entities for source:

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Campaign	Yes	Yes	No	Yes	Yes
Ad Set	Yes	Yes	No	Yes	Yes
Ads	Yes	Yes	No	Yes	Yes
Ad Creative	No	Yes	No	Yes	No
Insights - Account	No	Yes	No	Yes	No
Adaccounts	Yes	Yes	No	Yes	No
Insights - Ad	Yes	Yes	No	Yes	Yes
Insights - AdSet	Yes	Yes	No	Yes	Yes
Insights - Campaign	Yes	Yes	No	Yes	Yes

Example:

```
FacebookAds_read = glueContext.create_dynamic_frame.from_options(
    connection_type="FacebookAds",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "entityName",
        "API_VERSION": "v20.0"
    }
```

Facebook Ads entity and field details

For more information about the entities and field details see:

- Ad Account
- Campaign
- Ad Set
- Ad
- Ad Creative
- Insight Ad Account
- Insights Ads
- Insights AdSets
- Insights Campaigns

For more information, see Marketing API.

> **Note**
>
> Struct and List data types are converted to String data type in the response of the connectors.

Partitioning queries

You can provide the additional Spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, and `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition the query.
- `LOWER_BOUND`: an **inclusive** lower bound value of the chosen partition field.

  For the DateTime field, we accept the Spark timestamp format used in Spark SQL queries.

  Example of valid value:

  ```
  "2022-01-01"
  ```
- `UPPER_BOUND`: an **exclusive** upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: the number of partitions.

Example:

```
FacebookADs_read = glueContext.create_dynamic_frame.from_options(
    connection_type="FacebookAds",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "entityName",
        "API_VERSION": "v20.0",
        "PARTITION_FIELD": "created_time"
        "LOWER_BOUND": "2022-01-01"
        "UPPER_BOUND": "2024-01-02"
        "NUM_PARTITIONS": "10"
    }
```

Facebook Ads connection options

The following are connection options for Facebook Ads:

- `ENTITY_NAME`(String) - (Required) Used for read. The name of your object in Facebook Ads.
- `API_VERSION`(String) - (Required) Used for read. Facebook Ads Rest API version you want to use. For example: v1.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for read. Full Spark SQL query.

- `PARTITION_FIELD`(String) - Used for read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for read. Number of partitions for read.
- `TRANSFER_MODE`(String) - Default: SYNC. Used for asynchronous read.

Limitations and notes for Facebook Ads connector

The following are limitations or notes for the Facebook Ads connector:

- As Facebook Ads supports dynamic metadata, all fields can be queried. All the fields support filtration and records are fetched if the data is available, or else Facebook returns a Bad request (400) response with a proper error message.
- An app's call count is the number of calls a user can make during a rolling one-hour window 200 multiplied by the number of users. For rate limit details, see Rate Limits, and Business Use Case Rate Limits.

Connecting to Facebook Page Insights

Facebook Pages allow companies and other interest groups to create pages for the Facebook.com social network. Companies use these pages to share open hours, make announcements, and engage with customers online. If you are a Facebook Page Insights user, you can connect AWS Glue to your Facebook Page Insights account. You can use Facebook Page Insights as a data source in your ETL jobs. Run these jobs to transfer data from Facebook Page Insights to AWS services or other supported applications.

Topics

- AWS Glue support for Facebook Page Insights
- Policies containing the API operations for creating and using connections
- Configuring Facebook Page Insights
- Configuring Facebook Page Insights connections
- Reading from Facebook Page Insights entities
- Facebook Page Insights connection options

- Limitations and notes for Facebook Page Insights connector

AWS Glue support for Facebook Page Insights

AWS Glue supports Facebook Page Insights as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Facebook Page Insights.

Supported as a target?

No.

Supported Facebook Page Insights API versions

The following Facebook Page Insights API versions are supported:

- v17
- v18
- v19
- v20
- v21

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
          "glue:ListEntities",
          "glue:DescribeEntity"
        ],
        "Resource": "*"
      }
    ]
}
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Facebook Page Insights

Before you can use AWS Glue to transfer data from Facebook Page Insights, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- Facebook Standard accounts are accessed directly through Facebook.
- User authentication is needed to generate the access token.
- The Facebook Page Insights connector implements the User Access Token OAuth flow.
- The connector uses OAuth2.0 to authenticate our API requests to Facebook Page Insights. This falls under Multi-Factor Authentication (MFA) architecture, which is a superset of 2FA. It is web-based authentication.
- User needs to grant permissions to access the endpoints. For accessing the user's data, endpoint authorization is handled through permissions and features.

Configuring Facebook Page Insights connections

To configure a Facebook Page Insights connection:

1. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
    a. When selecting a **Connection type**, select Facebook Page Insights.
    b. Select the AWS IAM role which AWS Glue can assume and has permissions for following actions:

       JSON

       ```
       {
         "Version": "2012-10-17",
         "Statement": [
           {
             "Effect": "Allow",
             "Action": [
               "secretsmanager:DescribeSecret",
               "secretsmanager:GetSecretValue",
               "secretsmanager:PutSecretValue",
               "ec2:CreateNetworkInterface",
               "ec2:DescribeNetworkInterfaces",
               "ec2:DeleteNetworkInterface"
             ],
             "Resource": "*"
           }
         ]
       }
       ```
    c. Select the authorization code URL.
    d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.
    e. Select the network options if you want to use your network.
2. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Facebook Page Insights entities

Prerequisite

A Facebook Page Insights object you would like to read from. You will need the object name.

Supported entities for source:

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Page Content	Yes	No	Yes	Yes	Yes
Page CTA Clicks	Yes	No	No	Yes	Yes
Page Engagement	Yes	No	No	Yes	Yes
Page Impressions	Yes	No	No	Yes	Yes
Page Posts	Yes	No	No	Yes	Yes
Page Post Engagement	No	No	No	Yes	No
Page Post Reactions	No	No	No	Yes	No
Page Reactions	Yes	No	No	Yes	Yes
Stories	Yes	No	No	Yes	Yes
Page User Demographics	Yes	No	No	Yes	Yes
Page Video Views	Yes	No	No	Yes	Yes
Page Views	Yes	No	No	Yes	Yes

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Page Video Posts	Yes	No	No	Yes	Yes
Pages	No	Yes	No	Yes	No
Feeds	Yes	Yes	No	Yes	Yes

Example:

```
facebookPageInsights_read = glueContext.create_dynamic_frame. from options(
    connection_type="facebookpageinsights",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "entityName",
        "API_VERSION": "v21"
   }
```

Facebook Page Insights field details:

Entity	Field	Data type	Supported operators
Page Content	Name	String	N/A
	Period	Period	EQUAL_TO
	Since	DateTime	EQUAL_TO
	Values	List	N/A
	Title	String	N/A
	Description	String	N/A
	description_from_api_doc	String	N/A
	Id	String	N/A

Entity	Field	Data type	Supported operators
Page CTA Clicks	Name	String	N/A
	Period	Period	EQUAL_TO
	Since	DateTime	EQUAL_TO
	Values	List	N/A
	Title	String	N/A
	Description	String	N/A
	description_from_api_doc	String	N/A
	Id	String	N/A
Page Engagement	Name	String	N/A
	Period	Period	EQUAL_TO
	Since	DateTime	EQUAL_TO
	Values	List	N/A
	Title	String	N/A
	Description	String	N/A
	description_from_api_doc	String	N/A
	Id	String	N/A
Page Impressions	Name	String	N/A
	Period	Period	EQUAL_TO
	Since	DateTime	EQUAL_TO

Entity	Field	Data type	Supported operators
	Values	List	N/A
	Title	String	N/A
	Description	String	N/A
	description_from_api_doc	String	N/A
	Id	String	N/A
Page Posts	Name	String	N/A
	Period	Period	EQUAL_TO
	Since	DateTime	EQUAL_TO
	Values	List	N/A
	Title	String	N/A
	Description	String	N/A
	description_from_api_doc	String	N/A
	Id	String	N/A
Page Post Engagement	Name	String	N/A
	Period	Period	EQUAL_TO
	Values	List	N/A
	Title	String	N/A
	Description	String	N/A
	description_from_api_doc	String	N/A

Entity	Field	Data type	Supported operators
	Id	String	N/A
Page Post Reactions	Name	String	N/A
	Period	Period	EQUAL_TO
	Values	List	N/A
	Title	String	N/A
	Description	String	N/A
	description_from_api_doc	String	N/A
	Id	String	N/A
Page User Demographics	Name	String	N/A
	Period	Period	EQUAL_TO
	Since	DateTime	EQUAL_TO
	Values	List	N/A
	Title	String	N/A
	Description	String	N/A
	description_from_api_doc	String	N/A
	Id	String	N/A
Page Video Views	Name	String	N/A
	Period	Period	EQUAL_TO
	Since	DateTime	EQUAL_TO

Entity	Field	Data type	Supported operators
	Values	List	N/A
	Title	String	N/A
	Description	String	N/A
	description_from_api_doc	String	N/A
	Id	String	N/A
Page Views	Name	String	N/A
	Period	Period	EQUAL_TO
	Since	DateTime	EQUAL_TO
	Values	List	N/A
	Title	String	N/A
	Description	String	N/A
	description_from_api_doc	String	N/A
	Id	String	N/A
Page Video Posts	Name	String	N/A
	Period	Period	EQUAL_TO
	Since	DateTime	EQUAL_TO
	Values	List	N/A
	Title	String	N/A
	Description	String	N/A

Entity	Field	Data type	Supported operators
	description_from_api_doc	String	N/A
	Id	String	N/A
Pages	Name	String	N/A
	About	String	N/A
	access_token	String	N/A
	ad_campaign	String	N/A
	Affiliation	String	N/A
	app_id	String	N/A
	artists_we_like	String	N/A
	Attire	String	N/A
	Awards	String	N/A
	band_interests	String	N/A
	band_members	String	N/A
	best_page	String	N/A
	Bio	String	N/A
	Birthday	String	N/A
	booking_agent	String	N/A
	Built	String	N/A
	can_checkin	String	N/A
	can_post	String	N/A

Entity	Field	Data type	Supported operators
	Category	String	N/A
	category_list	List	N/A
	Checkins	Integer	N/A
	company_overview	String	N/A
	connected_instagram_account	String	N/A
	contact_address	String	N/A
	country_page_likes	Integer	N/A
	Cover	Struct	N/A
	culinary_team	String	N/A
	current_location	String	N/A
	delivery_and_pickup_option_info	List	N/A
	Description	String	N/A
	description_html	String	N/A
	differently_open_offerings	List	N/A
	directed_by	String	N/A
	display_subtext	String	N/A
	displayed_message_response_time	String	N/A
	Emails	String	N/A

Entity	Field	Data type	Supported operators
	Engagement	String	N/A
	fan_count	Integer	N/A
	featured_video	String	N/A
	Features	String	N/A
	followers_count	Integer	N/A
	food_styles	List	N/A
	Founded	String	N/A
	general_info	String	N/A
	general_manager	String	N/A
	Genre	String	N/A
	global_brand_page_name	String	N/A
	global_brand_root_id	String	N/A
	has_added_app	Boolean	N/A
	has_transitioned_to_new_page_experience	Boolean	N/A
	has_whatsapp_business_number	Boolean	N/A
	has_whatsapp_number	Boolean	N/A
	Hometown	String	N/A

Entity	Field	Data type	Supported operators
	Hours	Struct	N/A
	Impressum	String	N/A
	Influences	String	N/A
	instagram_business_account	String	N/A
	is_always_open	Boolean	N/A
	is_chain	Boolean	N/A
	is_community_page	Boolean	N/A
	is_eligible_for_branded_content	Boolean	N/A
	is_messenger_bot_get_started_enabled	Boolean	N/A
	is_messenger_platform_bot	Boolean	N/A
	is_owned	Boolean	N/A
	is_permanently_closed	Boolean	N/A
	is_published	Boolean	N/A
	Name	String	N/A
	Tasks	List	N/A
	is_unclaimed	Boolean	N/A
	is_webhooks_subscribed	Boolean	N/A

Entity	Field	Data type	Supported operators
	leadgen_tos_acceptance_time	DateTime	N/A
	leadgen_tos_accepted	Boolean	N/A
	leadgen_tos_accepting_user	String	N/A
	leadgen_tos_accepting_user	Struct	N/A
	Link	Link	N/A
	Location	Struct	N/A
	Members	String	N/A
	merchant_review_status	String	N/A
	messenger_ads_default_icebreakers	List	N/A
	messenger_ads_default_page_welcome_message	Struct	N/A
	messenger_ads_default_quick_replies	List	N/A
	messenger_ads_quick_replies_type	String	N/A
	Mission	String	N/A
	Mpg	String	N/A

Entity	Field	Data type	Supported operators
	name_with_location_descriptor	String	N/A
	Network	String	N/A
	new_like_count	Integer	N/A
	offer_eligible	Boolean	N/A
	overall_star_rating	Float	N/A
	page_token	String	N/A
	parent_page	String	N/A
	Parking	String	N/A
	payment_options	Struct	N/A
	personal_info	String	N/A
	personal_interests	String	N/A
	pharma_safety_info	String	N/A
	Phone	String	N/A
	pickup_options	List	N/A
	place_type	String	N/A
	plot_outline	String	N/A
	press_contact	String	N/A
	price_range	String	N/A
	privacy_info_url	String	N/A
	produced_by	String	N/A

Entity	Field	Data type	Supported operators
	Products	String	N/A
	promotion_eligible	Boolean	N/A
	promotion_ineligible_reason	String	N/A
	public_transit	String	N/A
	rating_count	Integer	N/A
	record_label	String	N/A
	release_date	String	N/A
	restaurant_services	Struct	N/A
	restaurant_specialties	Struct	N/A
	Schedule	String	N/A
	screenplay_by	String	N/A
	Season	String	N/A
	single_line_address	String	N/A
	Starring	String	N/A
	start_info	Struct	N/A
	store_code	String	N/A
	store_location_descriptor	String	N/A
	store_number	Integer	N/A
	Studio	String	N/A

Entity	Field	Data type	Supported operators
	supports_donate_button_in_live_video	Boolean	N/A
	talking_about_count	Integer	N/A
	temporary_status	String	N/A
	unread_message_count	Integer	N/A
	unread_notif_count	Integer	N/A
	unseen_message_count	Integer	N/A
	Username	String	N/A
	verification_status	String	N/A
	voip_info	Struct	N/A
	Website	String	N/A
	were_here_count	Integer	N/A
	whatsapp_number	String	N/A
written_by	String	N/A	
Feeds	Id	String	N/A
	Actions	List	N/A
	admin_creator	Object	N/A
	Application	Object	N/A
	Attachments	Objects	N/A
	backdated_time	DateTime	N/A

Entity	Field	Data type	Supported operators
	call_to_action	Object	N/A
	can_reply_privately	Boolean	N/A
	child_attachments	List	N/A
	Coordinates	Struct	N/A
	created_time	DateTime	N/A
	Event	Struct	N/A
	expanded_height	Integer	N/A
	expanded_width	Integer	N/A
	feed_targeting	Object	N/A
	From	Object	N/A
	full_picture	String	N/A
	Height	Integer	N/A
	Icon	String	N/A
	instagram_eligibility	String	N/A
	is_eligible_for_promotion	Boolean	N/A
	is_expired	Boolean	N/A
	is_hidden	Boolean	N/A
	is_inline_created	Boolean	N/A
	is_instagram_eligible	Boolean	N/A
	is_popular	Boolean	N/A

Entity	Field	Data type	Supported operators
	is_published	Boolean	N/A
	is_spherical	Boolean	N/A
	Message	String	N/A
	message_tags	List	N/A
	multi_share_end_card	Boolean	N/A
	multi_share_optimized	Boolean	N/A
	parent_id	String	N/A
	permalink_url	String	N/A
	Place	String	N/A
	Privacy	Object	N/A
	promotable_id	String	N/A
	promotion_status	String	N/A
	Properties	List	N/A
	scheduled_publish_time	Float	N/A
	Shares	Object	N/A
	status_type	String	N/A
	Story	String	N/A
	story_tags	List	N/A
	Subscribed	Boolean	N/A

Entity	Field	Data type	Supported operators
	Target	Struct	N/A
	Targeting	Object	N/A
	To	Object	N/A
	timeline_visibility	String	N/A
	updated_time	DateTime	N/A
	Via	Struct	N/A
	video_buying_eligibility	List	N/A
	Width	Integer	N/A
	Since	DateTime	EQUAL_TO

Partitioning queries

Filter-based partitioning:

You can provide the additional Spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, and `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition the query.
- `LOWER_BOUND`: an **inclusive** lower bound value of the chosen partition field.

  For the Datetime field, we accept the Spark timestamp format used in Spark SQL queries.

  Examples of valid value:

  ```
  "2024-09-30T01:01:01.000Z"
  ```
- `UPPER_BOUND`: an **exclusive** upper bound value of the chosen partition field.

- `NUM_PARTITIONS`: the number of partitions.

Example:

```
facebookPageInsights_read = glueContext.create_dynamic_frame.from_options(
     connection_type="facebookpageinsights",
     connection_options={
         "connectionName": "connectionName",
         "ENTITY_NAME": "entityName",
         "API_VERSION": "v21",
         "PARTITION_FIELD": "created_Time"
         "LOWER_BOUND": "2024-10-27T07:00:00+0000"
         "UPPER_BOUND": "2024-10-27T07:00:00+0000"
         "NUM_PARTITIONS": "10"
     }
```

Facebook Page Insights connection options

The following are connection options for Facebook Page Insights:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your object in Facebook Page Insights.
- `API_VERSION`(String) - (Required) Used for Read. Facebook Page Insights Rest API version you want to use.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.
- `INSTANCE_URL`(String) - (Required) Used for Read. A valid Facebook Page Insights instance URL.

Limitations and notes for Facebook Page Insights connector

The following are limitations or notes for the Facebook Page Insights connector:

- Most metrics will update once every 24 hours.
- Only the last two years of insights data is available.
- Only 90 days of insights can be viewed at one time when using the `since` and `until` parameters.

Connecting to Freshdesk

Freshdesk is a cloud-based customer support software that is both feature-rich and easy to use. With multiple support channels available, including live chat, email, phone, and social media, you can help customers through their preferred communication method. If you are a Freshdesk user, you can connect AWS Glue to your Freshdesk account. You can use Freshdesk as a data source in your ETL jobs. Run these jobs to transfer data from Freshdesk to AWS services or other supported applications.

Topics

- AWS Glue support for Freshdesk
- Policies containing the API operations for creating and using connections
- Configuring Freshdesk
- Configuring Freshdesk connections
- Reading from Freshdesk entities
- Freshdesk connection options
- Limitations and notes for Freshdesk connector

AWS Glue support for Freshdesk

AWS Glue supports Freshdesk as follows:

Supported as a source?

Yes – Sync and Async. You can use AWS Glue ETL jobs to query data from Freshdesk.

Supported as a target?

No.

Supported Freshdesk API versions

The following Freshdesk API versions are supported:

- v2

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.

- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Freshdesk

Before you can use AWS Glue to transfer data from Freshdesk, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- A Freshdesk account. You can choose from Free, Growth, Pro or Enterprise editions.
- A Freshdesk user's API key.

Configuring Freshdesk connections

Freshdesk supports custom authentication.

For public Freshdesk documentation on generating the required API keys for custom authorization, see Freshdesk authentication.

The following are the steps to configure Freshdesk connection:

- In AWS Secrets Manager, create a secret with the following details:
  - For customer managed connected app – the secret should contain the connected app API key with `apiKey` as key. Note that you must create a secret per connection in AWS Glue.
- In the AWS Glue Studio, create a connection under **Data Connections** by following the steps below:
  - When selecting a **Data source**, select Freshdesk.
  - Provide the `INSTANCE_URL` of the Freshdesk instance you want to connect to.
  - Select the AWS IAM role which AWS Glue can assume and has permissions for following actions:

    JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

  - Select the `secretName` that you want to use for this connection in AWS Glue to put the tokens.
  - Select the network options if you want to use your network.
- Grant the IAM role associated with your AWS Glue job permission to read `secretName`.
- In your AWS Glue job configuration, provide `connectionName` as an **Additional network connection**.

Reading from Freshdesk entities

Prerequisite

A Freshdesk object you would like to read from. You will need the object name.

Supported entities for Sync source:

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Agents	Yes	Yes	No	Yes	Yes
Business Hours	No	Yes	No	Yes	Yes
Company	Yes	Yes	No	Yes	Yes

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Contacts	Yes	Yes	No	Yes	Yes
Conversations	No	Yes	No	Yes	No
Email Configs	No	Yes	No	Yes	No
Email Inboxes	Yes	Yes	Yes	Yes	No
Forum Categories	No	Yes	No	Yes	No
Forums	No	Yes	No	Yes	No
Groups	No	Yes	No	Yes	No
Products	No	Yes	No	Yes	No
Roles	No	Yes	No	Yes	No
Satisfaction Ratings	Yes	Yes	No	Yes	No
Skills	No	Yes	No	Yes	No
Solutions	Yes	Yes	No	Yes	No
Surveys	No	Yes	No	Yes	No
Tickets	Yes	Yes	Yes	Yes	Yes
Time Entries	Yes	Yes	No	Yes	No
Topics	No	Yes	No	Yes	No
Topic Comments	No	Yes	No	Yes	No

Supported entities for Async source:

Entity	API version	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Companies	v2	No	No	No	No	No
Contacts	v2	No	No	No	No	No

Example:

```
freshdesk_read = glueContext.create_dynamic_frame.from_options(
    connection_type="freshdesk",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "entityName",
        "API_VERSION": "v2"
    }
```

Freshdesk entity and field details:

Entity	Field
Agents	https://developers.freshdesk.com/api/#list_all_agents
Business-hours	https://developers.freshdesk.com/api/#list_all_business_hours
Comments	https://developers.freshdesk.com/api/#comment_attributess
Company	https://developers.freshdesk.com/api/#companies
Contacts	https://developers.freshdesk.com/api/#list_all_contacts

Entity	Field
Conversations	https://developers.freshdesk.com/api/#list_all_ticket_notes
Email-configs	https://developers.freshdesk.com/api/#list_all_email_configs
Email-inboxes	https://developers.freshdesk.com/api/#list_all_email_mailboxes
Forum-categories	https://developers.freshdesk.com/api/#category_attributes
Forums	https://developers.freshdesk.com/api/#forum_attributes
Groups	https://developers.freshdesk.com/api/#list_all_groups
Products	https://developers.freshdesk.com/api/#list_all_products
Roles	https://developers.freshdesk.com/api/#list_all_roles
Satisfaction-rating	https://developers.freshdesk.com/api/#view_all_satisfaction_ratingss
Skills	https://developers.freshdesk.com/api/#list_all_skills
Solutions	https://developers.freshdesk.com/api/#solution_content
Surveys	https://developers.freshdesk.com/api/#list_all_survey
Tickets	https://developers.freshdesk.com/api/#list_all_tickets

Entity	Field
Time-entries	https://developers.freshdesk.com/api/#list_all_time_entries
Topics	https://developers.freshdesk.com/api/#topic_attributes

Partitioning queries

Filter-based partitioning:

You can provide the additional Spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, and `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition the query.
- `LOWER_BOUND`: an **inclusive** lower bound value of the chosen partition field.

  For the Datetime field, we accept the Spark timestamp format used in Spark SQL queries.

  Examples of valid value:

  ```
  "2024-09-30T01:01:01.000Z"
  ```
- `UPPER_BOUND`: an **exclusive** upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: the number of partitions.

Example:

```
freshDesk_read = glueContext.create_dynamic_frame.from_options(
     connection_type="freshdesk",
     connection_options={
         "connectionName": "connectionName",
         "ENTITY_NAME": "entityName",
         "API_VERSION": "v2",
         "PARTITION_FIELD": "Created_Time"
         "LOWER_BOUND": " 2024-10-27T23:16:08Z“
```

```
        "UPPER_BOUND": " 2024-10-27T23:16:08Z"
        "NUM_PARTITIONS": "10"
    }
```

Freshdesk connection options

The following are connection options for Freshdesk:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your object in Freshdesk.
- `API_VERSION`(String) - (Required) Used for Read. Freshdesk Rest API version you want to use.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.
- `INSTANCE_URL`(String) - (Required) Used for Read. A valid Freshdesk instance URL.
- `TRANSFER_MODE`(String) - Used to indicate whether the type of processing like `SYNC` or `ASYNC` is set to `SYNC` by default. (Optional)

Limitations and notes for Freshdesk connector

The following are limitations or notes for the Freshdesk connector:

- The `Company`, `Contacts`, and `Tickets` entities with filtration have pagination limitations. They return only 30 records per page and the page value can be set up to a maximum of 10 (fetching a maximum of 300 records).
- The `Tickets` entity does not fetch records older than 30 days.
- The `Company`, `Contacts`, and `Tickets` entities support the 'Date' datatype in filtration. You should select the 'Daily' onward trigger frequencies for these three entities. Selecting 'Minutes' or

'Hourly' can lead to duplicate data. Also, while selecting these fields for filtration, only the date value should be selected, since it will only consider the date portion of the selected timestamp.

- The number of API calls per minute is based on your plan. This limit is applied on an account wide basis irrespective of factors such as the number of agents or IP addresses used to make the calls. For all trial users, there is a default API limit of 50 calls/minute. For more details, refer to Freshdesk
- For any entity, only one Export/Async Job is processed at a time. A new job will only be processed once the existing job has completed successfully or failed. For more details, refer to Freshdesk
- The following fields are supported for Sync API calls, but are not supported/allowed to be passed in Async API request body.
  - id
  - created_at
  - updated_at
  - updated_since
  - active
  - company_id
  - other_companies
  - avatar
  - view_all_tickets
  - deleted
  - other_emails
  - state
  - tag
  - tags

Connecting to Freshsales

Freshsales is an intuitive CRM that helps sales reps take the guesswork out of sales. With the built-in phone and email, tasks, appointments and notes, sales reps needn't have to toggle between tabs to follow up on prospects. You can manage your deals better with the pipeline view and drive more deals to closure. If you are a Freshsales user, you can connect AWS Glue to your Freshsales account. You can use Freshsales as a data source in your ETL jobs. Run these jobs to transfer data from Freshsales to AWS services or other supported applications.

Topics

- AWS Glue support for Freshsales
- Policies containing the API operations for creating and using connections
- Configuring Freshsales
- Configuring Freshsales connections
- Reading from Freshsales entities
- Freshsales connection options
- Freshsales limitations

AWS Glue support for Freshsales

AWS Glue supports Freshsales as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Freshsales.

Supported as a target?

No.

Supported Freshsales API versions

The following Freshsales API versions are supported:

- v1.0

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Freshsales

Before you can use AWS Glue to transfer data from Freshsales, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- You have a Freshsales account.
- You have a user API key.

If you meet these requirements, you're ready to connect AWS Glue to your Freshsales account. For typical connections, you don't need do anything else in Freshsales.

Configuring Freshsales connections

Freshsales supports custom authentication.

For public Freshsales documentation on generating the required API keys for custom authentication, see [Authentication](#).

To configure a Freshsales connection:

1. In AWS Secrets Manager, create a secret with the following details:
   a. For the customer managed connected app, the Secret should contain the connected app API key with *`apiSecretKey`* as key. The Secret also needs to contain another key-value pair with *`apiKey`* as key and *`token`* as value.
   b. Note: you must create a secret for your connections in AWS Glue.

1. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Data Source**, select Freshsales.
   b. Provide the `INSTANCE_URL` of the Freshsales account you want to connect to.
   c. Select the AWS IAM role which AWS Glue can assume and has permissions for following actions:

      JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
        "Resource": "*"
      }
    ]
  }
```

   d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.
   e. Select the network options if you want to use your network.
2. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Freshsales entities

Prerequisite

A Freshsales object you would like to read from. You will need the object name.

Supported entities for source:

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Accounts	Yes	Yes	Yes	Yes	Yes
Contacts	Yes	Yes	Yes	Yes	Yes

Example:

```
freshSales_read = glueContext.create_dynamic_frame.from_options(
     connection_type="freshsales",
     connection_options={
         "connectionName": "connectionName",
         "ENTITY_NAME": "entityName",
         "API_VERSION": "v1.0"
     }
```

Freshsales entity and field details:

Freshsales provides endpoints to fetch metadata dynamically for supported entities. Accordingly, operator support is captured at the datatype level.

Entity	Data type	Supported operators
Freshsale entities (all)	Integer	!=,=,<,<=,>,>=,BETWEEN
	String	Like, =, !=
	BigInteger	!=,=,<,<=,>,>=,BETWEEN
	Boolean	=
	Double	!=,=,<,<=,>,>=,BETWEEN
	BigDecimal	!=,=,<,<=,>,>=,BETWEEN
	Date	!=,=,<,<=,>,>=,BETWEEN
	DateTime	!=,=,<,<=,>,>=,BETWEEN
	Struct	N/A
	List	N/A

Partitioning queries

Filter-based partitioning:

You can provide the additional Spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, and `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition the query.
- `LOWER_BOUND`: an **inclusive** lower bound value of the chosen partition field.

  For the Datetime field, we accept the value in ISO format.

  Examples of valid value:

  ```
  "2024-09-30T01:01:01.000Z"
  ```

- `UPPER_BOUND`: an **exclusive** upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: the number of partitions.

Example:

```
freshSales_read = glueContext.create_dynamic_frame.from_options(
     connection_type="freshsales",
     connection_options={
         "connectionName": "connectionName",
         "ENTITY_NAME": "entityName",
         "API_VERSION": "v1",
         "PARTITION_FIELD": "Created_Time"
         "LOWER_BOUND": " 2024-10-15T21:16:25Z"
         "UPPER_BOUND": " 2024-10-20T21:25:50Z"
         "NUM_PARTITIONS": "10"
     }
```

Freshsales connection options

The following are connection options for Freshsales:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your object in Freshsales.
- `API_VERSION`(String) - (Required) Used for Read. Freshsales Rest API version you want to use.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.
- `INSTANCE_URL`(String) - Used for Read. A valid Freshsales instance URL.

Freshsales limitations

The following are limitations or notes for Freshsales:

- In Freshsales, the API Rate limit is 1000 API requests per hour per account (see Errors. But this limit is extendable with the Enterprise subscription plan (see the plan comparison).

Connecting to Google Ads

The Google Ads API is the programmatic interface to Google Ads, used for managing large or complex Google Ads accounts and campaigns. If you're a Google Ads user, you can connect AWS Glue to your Google Ads account. Then, you can use Google Ads as a data source in your ETL jobs. Run these jobs to transfer data between Google Ads and AWS services or other supported applications.

Topics

- AWS Glue support for Google Ads
- Policies containing the API operations for creating and using connections
- Configuring Google Ads
- Configuring Google Ads connections
- Reading from Google Ads entities
- Google Ads connection options
- Creating a Google Ads account
- Limitations

AWS Glue support for Google Ads

AWS Glue supports Google Ads as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Google Ads.

Supported as a target?

No.

Supported Google Ads API versions

v18

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

You can also use the following managed IAM policies to allow access:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Google Ads

Before you can use AWS Glue to transfer from Google Ads, you must meet these requirements:

Minimum requirements

- You have a Google Ads account with Email and Password. For more information on creating an account, see Creating a Google Ads account.
- Your Google Ads account is enabled for API access. All use of the Google Ads API is available at no additional cost.
- Your Google Ads account allows you to install connected apps. If you lack access to this functionality, contact your Google Ads administrator.

If you meet these requirements, you’re ready to connect AWS Glue to your Google Ads account.

Configuring Google Ads connections

Google Ads supports `AUTHORIZATION_CODE` grant type for OAuth2.

This grant type is considered “three-legged” OAuth as it relies on redirecting users to the third party authorization server to authenticate the user. It is used when creating connections via the AWS Glue Console. The AWS Glue Console will redirect the user to Google Ads where the user must login and allow AWS Glue the requested permissions to access their Google Ads instance.

Users may opt to create their own connected app in Google Ads and provide their own client ID and client secret when creating connections through the AWS Glue Console. In this scenario, they will still be redirected to Google Ads to login and authorize AWS Glue to access their resources.

This grant type results in a refresh token and access token. The access token is short lived, and may be refreshed automatically without user interaction using the refresh token.

For more information, see public Google Ads documentation on creating a connected app for Authorization Code OAuth flow .

To configure a Google Ads connection:

1. In AWS Secrets Manager, create a secret with the following details. It is required to create a secret for each connection in AWS Glue.
    a. For AuthorizationCode grant type:

   - For customer managed connected app – Secret should contain the connected app Consumer Secret with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as key.

2. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Connection type**, select Google Ads.
   b. Provide the `developer token` of the Google Ads you want to connect to.
   c. Provide the `MANAGER ID` of the Google Ads if you want to log in as manager.
   d. Select the IAM role which AWS Glue can assume and has permissions for following actions:

      JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

   e. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.
   f. Select the network options if you want to use your network.
3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Google Ads entities

Prerequisites

- A Google Ads Object you would like to read from. Refer the supported entities table below to check the available entities.

Supported entities

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Ad Group Ad	Yes	Yes	Yes	No	Yes
Ad Group	Yes	Yes	Yes	No	Yes
Campaign Budget	Yes	Yes	Yes	Yes	Yes
Account Budget	Yes	No	Yes	Yes	No
Campaign	Yes	Yes	Yes	Yes	Yes
Account	Yes	No	Yes	No	No

Example

```
googleAds_read = glueContext.create_dynamic_frame.from_options(
    connection_type="googleads",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "campaign-3467***",
        "API_VERSION": "v16"
    }
```

Google Ads entity and field details

Entity	Field	Data Type	Supported Operators
Account	resourceName	String	!=, =

Entity	Field	Data Type	Supported Operators
Account	callReportingEnabled	Boolean	!=, =
Account	callConversionReportingEnabled	Boolean	!=, =
Account	callConversionAction	String	!=, =
Account	conversionTrackingId	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Account	crossAccountConversionTrackingId	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Account	payPerConversionEligibilityFailureReasons	List	
Account	id	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Account	currencyCode	String	!=, =, LIKE
Account	timeZone	String	!=, =, LIKE
Account	autoTaggingEnabled	Boolean	!=, =
Account	hasPartnersBadge	Boolean	!=, =
Account	manager	Boolean	!=, =
Account	testAccount	Boolean	!=, =
Account	date	Date	BETWEEN, =, <, >, <=, >=
Account	costMicros	BigInteger	BETWEEN, =, !=, <, >, <=, >=

Entity	Field	Data Type	Supported Operators
Account	acceptedCustomerDataTerms	Boolean	
Account	conversionTrackingStatus	String	!=, =, LIKE
Account	enhancedConversionsForLeadsEnabled	Boolean	
Account	googleAdsConversionCustomer	String	
Account	status	String	!=, =
Account	allConversionsByConversionDate	Double	!=, =, <, >
Account	allConversionsValueByConversionDate	Double	!=, =, <, >
Account	conversionsByConversionDate	Double	!=, =, <, >
Account	conversionsValueByConversionDate	Double	!=, =, <, >
Account	valuePerAllConversionsByConversionDate	Double	!=, =, <, >
Account	videoViews	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Account	clicks	BigInteger	BETWEEN, =, !=, <, >, <=, >=

Entity	Field	Data Type	Supported Operators
Account	invalidClicks	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Account	costPerAllConversions	Double	!=, =, <, >
Account	costPerConversion	Double	!=, =, <, >
Account	conversions	Double	!=, =, <, >
Account	absoluteTopImpressionPercentage	Double	!=, =, <, >
Account	impressions	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Account	topImpressionPercentage	Double	!=, =, <, >
Account	averageCpc	Double	!=, =, <, >
Account	activeViewMeasurableCostMicros	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Account	averageCost	Double	!=, =, <, >
Account	ctr	Double	!=, =, <, >
Account	activeViewCtr	Double	!=, =, <, >
Account	searchImpressionShare	Double	!=, =, <, >
Account	conversionAction	String	!=, =
Account	conversionActionCategory	String	!=, =

Entity	Field	Data Type	Supported Operators
Account	conversionActionName	String	!=, =, LIKE
Account Budget	resourceName	String	!=, =
Account Budget	status	String	!=, =
Account Budget	proposedEndTimeType	String	!=, =
Account Budget	approvedEndTimeType	String	!=, =
Account Budget	id	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Account Budget	billingSetup	String	!=, =
Account Budget	name	String	!=, =, LIKE
Account Budget	approvedStartDateTime	DateTime	BETWEEN, =, <, >, <=, >=
Account Budget	proposedSpendingLimitMicros	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Account Budget	approvedSpendingLimitMicros	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Account Budget	adjustedSpendingLimitMicros	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Account Budget	amountServedMicros	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Ad Group	resourceName	String	!=, =, LIKE

Entity	Field	Data Type	Supported Operators
Ad Group	status	String	!=, =, LIKE
Ad Group	type	String	!=, =, LIKE
Ad Group	id	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Ad Group	name	String	!=, =, LIKE
Ad Group	campaign	String	!=, =
Ad Group	cpcBidMicros	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Ad Group	targetCpaMicros	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Ad Group	cpmBidMicros	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Ad Group	cpvBidMicros	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Ad Group	targetCpmMicros	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Ad Group	effectiveTargetCpaMicros	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Ad Group	date	Date	BETWEEN, =, <, >, <=, >=
Ad Group	costMicros	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Ad Group	useAudienceGrouped	Boolean	!=, =

Entity	Field	Data Type	Supported Operators
Ad Group	effectiveCpcBidMicros	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Ad Group	allConversionsByConversionDate	Double	!=, =, <, >
Ad Group	allConversionsValueByConversionDate	Double	!=, =, <, >
Ad Group	conversionsByConversionDate	Double	!=, =, <, >
Ad Group	conversionsValueByConversionDate	Double	!=, =, <, >
Ad Group	valuePerAllConversionsByConversionDate	Double	!=, =, <, >
Ad Group	valuePerConversionsByConversionDate	Double	!=, =, <, >
Ad Group	averageCost	Double	!=, =, <, >
Ad Group	costPerAllConversions	Double	!=, =, <, >
Ad Group	costPerConversion	Double	!=, =, <, >
Ad Group	averagePageViews	Double	!=, =, <, >
Ad Group	videoViews	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Ad Group	clicks	BigInteger	BETWEEN, =, !=, <, >, <=, >=

Entity	Field	Data Type	Supported Operators
Ad Group	allConversions	Double	!=, =, <, >
Ad Group	averageCpc	Double	!=, =, <, >
Ad Group	absoluteTopImpressionPercentage	Double	!=, =, <, >
Ad Group	impressions	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Ad Group	topImpressionPercentage	Double	!=, =, <, >
Ad Group	activeViewCtr	Double	!=, =, <, >
Ad Group	ctr	Double	!=, =, <, >
Ad Group	searchTopImpressionShare	Double	!=, =, <, >
Ad Group	searchImpressionShare	Double	!=, =, <, >
Ad Group	searchAbsoluteTopImpressionShare	Double	!=, =, <, >
Ad Group	relativeCtr	Double	!=, =, <, >
Ad Group	conversionAction	String	!=, =
Ad Group	conversionActionCategory	String	!=, =
Ad Group	conversionActionName	String	!=, =, LIKE
Ad Group	updateMask	String	

Entity	Field	Data Type	Supported Operators
Ad Group	create	Struct	
Ad Group	update	Struct	
Ad Group	primaryStatus	String	!=, =
Ad Group	primaryStatusReasons	List	
Ad Group Ad	resourceName	String	!=, =
Ad Group Ad	id	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Ad Group Ad	status	String	!=, =
Ad Group Ad	labels	List	
Ad Group Ad	adGroup	String	!=, =
Ad Group Ad	costMicros	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Ad Group Ad	approvalStatus	String	!=, =
Ad Group Ad	reviewStatus	String	!=, =
Ad Group Ad	adStrength	String	!=, =
Ad Group Ad	type	String	!=, =
Ad Group Ad	businessName	String	!=, =, LIKE
Ad Group Ad	date	Date	BETWEEN, =, <, >, <=, >=
Ad Group Ad	allConversionsByConversionDate	Double	!=, =, <, >

Entity	Field	Data Type	Supported Operators
Ad Group Ad	allConversionsValueByConversionDate	Double	!=, =, <, >
Ad Group Ad	conversionsByConversionDate	Double	!=, =, <, >
Ad Group Ad	conversionsValueByConversionDate	Double	!=, =, <, >
Ad Group Ad	valuePerAllConversionsByConversionDate	Double	!=, =, <, >
Ad Group Ad	valuePerConversionsByConversionDate	Double	!=, =, <, >
Ad Group Ad	activeViewMeasurableCostMicros	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Ad Group Ad	averageCost	Double	!=, =, <, >
Ad Group Ad	costPerAllConversions	Double	!=, =, <, >
Ad Group Ad	costPerConversion	Double	!=, =, <, >
Ad Group Ad	clicks	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Ad Group Ad	averagePageViews	Double	!=, =, <, >
Ad Group Ad	videoViews	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Ad Group Ad	allConversions	Double	!=, =, <, >
Ad Group Ad	averageCpc	Double	!=, =, <, >

Entity	Field	Data Type	Supported Operators
Ad Group Ad	topImpressionPercentage	Double	!=, =, <, >
Ad Group Ad	impressions	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Ad Group Ad	absoluteTopImpressionPercentage	Double	!=, =, <, >
Ad Group Ad	activeViewCtr	Double	!=, =, <, >
Ad Group Ad	ctr	Double	!=, =, <, >
Ad Group Ad	conversionAction	String	!=, =
Ad Group Ad	conversionActionCategory	String	!=, =
Ad Group Ad	conversionActionName	String	!=, =, LIKE
Ad Group Ad	updateMask	String	
Ad Group Ad	create	Struct	
Ad Group Ad	update	Struct	
Ad Group Ad	policyValidationParameter	Struct	
Ad Group Ad	primaryStatus	String	!=, =
Ad Group Ad	primaryStatusReasons	List	
Campaign	resourceName	String	!=, =
Campaign	status	String	!=, =

Entity	Field	Data Type	Supported Operators
Campaign	baseCampaign	String	!=, =
Campaign	name	String	!=, =, LIKE
Campaign	id	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Campaign	campaignBudget	String	!=, =, LIKE
Campaign	startDate	Date	BETWEEN, =, <, >, <=, >=
Campaign	endDate	Date	BETWEEN, =, <, >, <=, >=
Campaign	adServingOptimizationStatus	String	!=, =
Campaign	advertisingChannelType	String	!=, =
Campaign	advertisingChannelSubType	String	!=, =
Campaign	experimentType	String	!=, =
Campaign	servingStatus	String	!=, =
Campaign	biddingStrategyType	String	!=, =
Campaign	domainName	String	!=, =, LIKE
Campaign	languageCode	String	!=, =, LIKE
Campaign	useSuppliedUrlsOnly	Boolean	!=, =
Campaign	positiveGeoTargetType	String	!=, =

Entity	Field	Data Type	Supported Operators
Campaign	negativeGeoTargetType	String	!=, =
Campaign	paymentMode	String	!=, =
Campaign	optimizationGoalTypes	List	
Campaign	date	Date	BETWEEN, =, <, >, <=, >=
Campaign	averageCost	Double	
Campaign	clicks	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Campaign	costMicros	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Campaign	impressions	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Campaign	useAudienceGrouped	Boolean	!=, =
Campaign	activeViewMeasurableCostMicros	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Campaign	costPerAllConversions	Double	!=, =, <, >
Campaign	costPerConversion	Double	!=, =, <, >
Campaign	invalidClicks	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Campaign	publisherPurchasedClicks	BigInteger	BETWEEN, =, !=, <, >, <=, >=

Entity	Field	Data Type	Supported Operators
Campaign	averagePageViews	Double	!=, =, <, >
Campaign	videoViews	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Campaign	allConversionsByConversionDate	Double	!=, =, <, >
Campaign	allConversionsValueByConversionDate	Double	!=, =, <, >
Campaign	conversionsByConversionDate	Double	!=, =, <, >
Campaign	conversionsValueByConversionDate	Double	!=, =, <, >
Campaign	valuePerAllConversionsByConversionDate	Double	!=, =, <, >
Campaign	valuePerConversionsByConversionDate	Double	!=, =, <, >
Campaign	allConversions	Double	!=, =, <, >
Campaign	absoluteTopImpressionPercentage	Double	!=, =, <, >
Campaign	searchAbsoluteTopImpressionShare	Double	!=, =, <, >
Campaign	averageCpc	Double	!=, =, <, >
Campaign	searchImpressionShare	Double	!=, =, <, >

Entity	Field	Data Type	Supported Operators
Campaign	searchTopImpressionShare	Double	!=, =, <, >
Campaign	activeViewCtr	Double	!=, =, <, >
Campaign	ctr	Double	!=, =, <, >
Campaign	relativeCtr	Double	!=, =, <, >
Campaign	updateMask	String	
Campaign	create	Struct	
Campaign	update	Struct	
Campaign Budget	resourceName	String	!=, =
Campaign Budget	id	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Campaign Budget	status	String	!=, =
Campaign Budget	deliveryMethod	String	!=, =
Campaign Budget	period	String	!=, =
Campaign Budget	type	String	!=, =
Campaign Budget	name	String	!=, =, LIKE
Campaign Budget	amountMicros	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Campaign Budget	explicitlyShared	Boolean	!=, =
Campaign Budget	referenceCount	BigInteger	BETWEEN, =, !=, <, >, <=, >=

Entity	Field	Data Type	Supported Operators
Campaign Budget	hasRecommendedBudget	Boolean	!=, =
Campaign Budget	date	Date	BETWEEN, =, <, >, <=, >=
Campaign Budget	costMicros	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Campaign Budget	startDate	Date	BETWEEN, =, <, >, <=, >=
Campaign Budget	endDate	Date	BETWEEN, =, <, >, <=, >=
Campaign Budget	maximizeConversionValueTargetRoas	Double	!=, =, <, >
Campaign Budget	maximizeConversionsTargetCpaMicros	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Campaign Budget	selectiveOptimizationConversionActions	String	
Campaign Budget	averageCost	Double	!=, =, <, >
Campaign Budget	costPerAllConversions	Double	!=, =, <, >
Campaign Budget	costPerConversion	Double	!=, =, <, >
Campaign Budget	videoViews	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Campaign Budget	clicks	BigInteger	BETWEEN, =, !=, <, >, <=, >=

Entity	Field	Data Type	Supported Operators
Campaign Budget	allConversions	Double	!=, =, <, >
Campaign Budget	valuePerAllConversions	Double	!=, =, <, >
Campaign Budget	averageCpc	Double	!=, =, <, >
Campaign Budget	impressions	BigInteger	BETWEEN, =, !=, <, >, <=, >=
Campaign Budget	ctr	Double	!=, =, <, >
Campaign Budget	updateMask	String	
Campaign Budget	create	Struct	
Campaign Budget	update	Struct	

Partitioning queries

Additional spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, `NUM_PARTITIONS` can be provided if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition query.
- `LOWER_BOUND`: an inclusive lower bound value of the chosen partition field.

  For date, we accept the Spark date format used in Spark SQL queries. Example of valid values: `"2024-02-06"`.
- `UPPER_BOUND`: an exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: number of partitions.

Entity-wise partitioning field support details are captured in the following table.

Entity Name	Partitioning Field	Data Type
Ad Group Ad	date	Date
Ad Group	date	Date
Campaign	date	Date
Campaign Budget	date	Date

Example

```
googleads_read = glueContext.create_dynamic_frame.from_options(
    connection_type="googleads",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "campaign-3467***",
        "API_VERSION": "v16",
        "PARTITION_FIELD": "date"
        "LOWER_BOUND": "2024-01-01"
        "UPPER_BOUND": "2024-06-05"
        "NUM_PARTITIONS": "10"
    }
)
```

Google Ads connection options

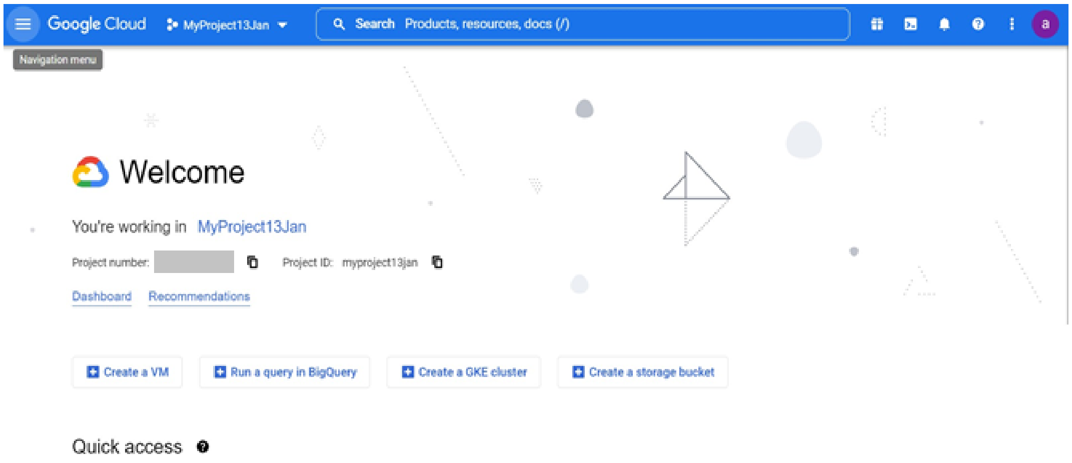
The following are connection options for Google Ads:

- `ENTITY_NAME`(String) - (Required) Used for Read/Write. The name of your Object in Google Ads.
- `API_VERSION`(String) - (Required) Used for Read/Write. Google Ads Rest API version you want to use. Example: v16.
- `DEVELOPER_TOKEN`(String) - (Required) Used for Read/Write. Required to authenticate the developer or application making requests to the API.
- `MANAGER_ID`(String) - Used for Read/Write. A unique identifier that allows you to manage multiple Google Ads accounts. This is the customer ID of the authorized manager. If your access to the customer account is through a manager account, the `MANAGER_ID` is required. For more information, see login-customer-id.

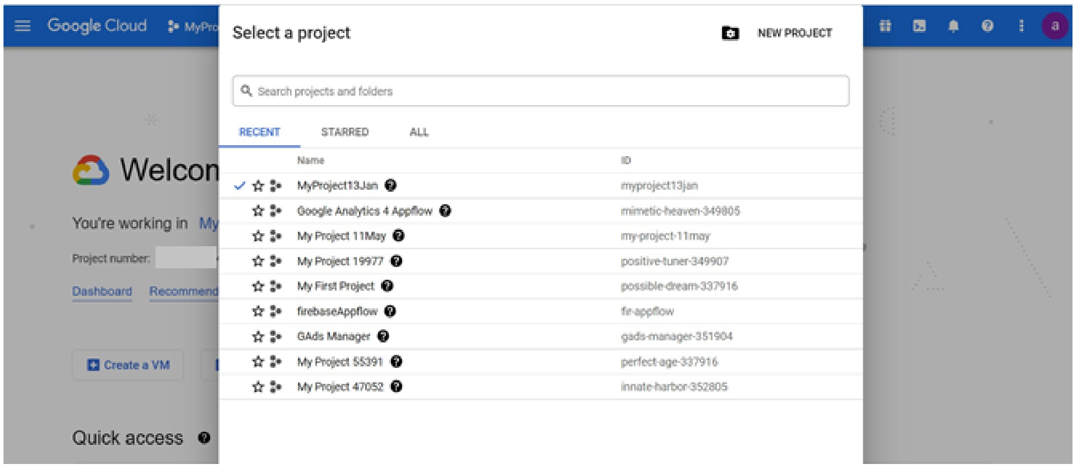
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.

Creating a Google Ads account

1. Log in to [Google Ads Developer Account]() with your credentials, and go to *MyProject.


Google Cloud
MyProject13Jan
Search Products, resources, docs (/)
a
Navigation menu
Welcome
You're working in MyProject13Jan
Project number: [redacted]
Project ID: myproject13jan
Dashboard
Recommendations
Create a VM
Run a query in BigQuery
Create a GKE cluster
Create a storage bucket
Quick access

2. Choose **New Project** and provide the information which is required for creating Google project if you don't have any registered application in it.



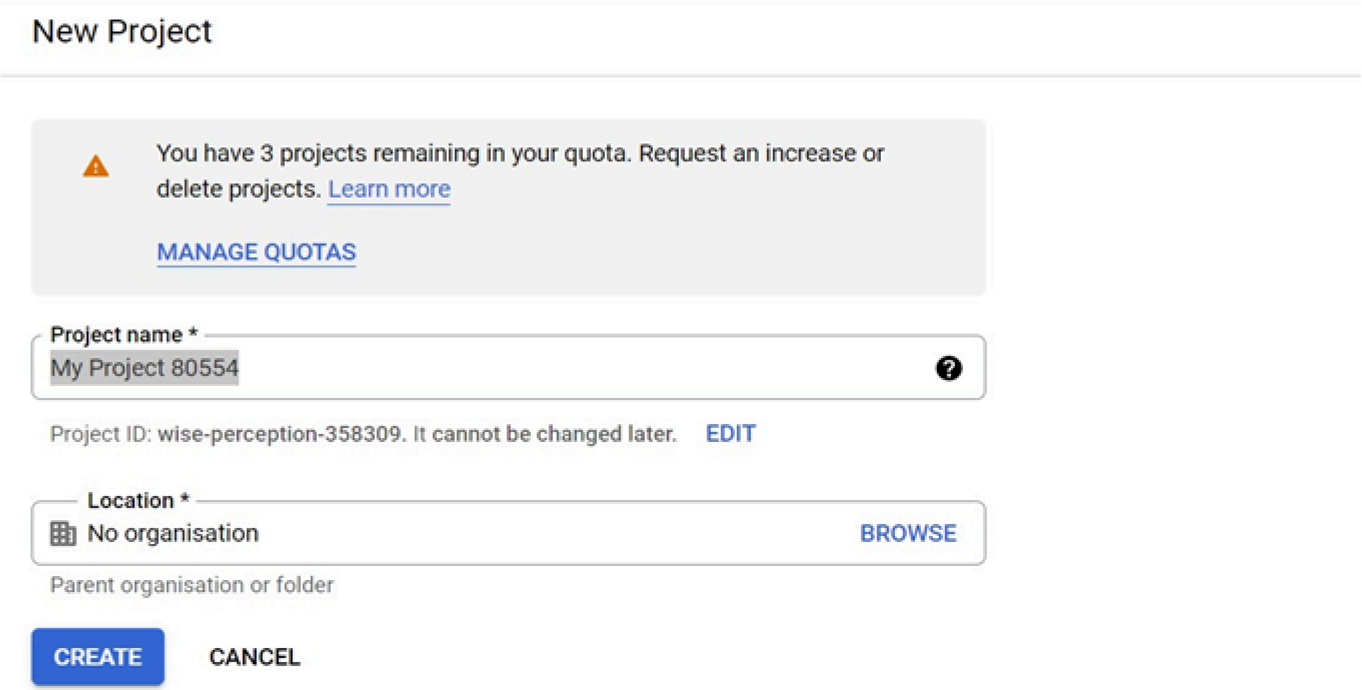
Select a project

NEW PROJECT

Search projects and folders

RECENT STARRED ALL

Name	ID
✓ MyProject13Jan	myproject13jan
Google Analytics 4 Appflow	mimetic-heaven-349805
My Project 11May	my-project-11may
My Project 19977	positive-tuner-349907
My First Project	possible-dream-337916
firebaseAppflow	fir-appflow
GAds Manager	gads-manager-351904
My Project 55391	perfect-age-337916
My Project 47052	innate-harbor-352805



New Project

You have 3 projects remaining in your quota. Request an increase or delete projects. Learn more

MANAGE QUOTAS

Project name * My Project 80554

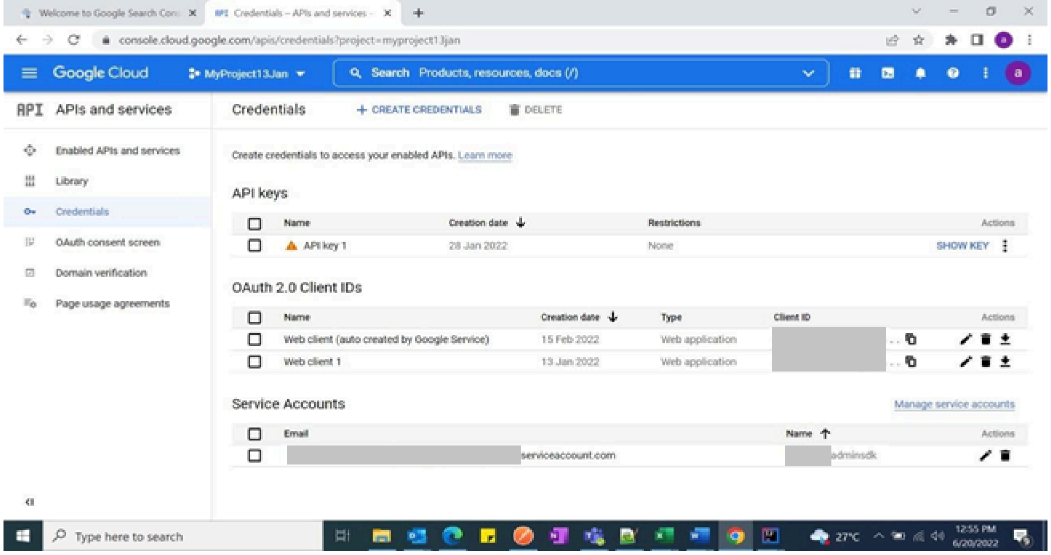
Project ID: wise-perception-358309. It cannot be changed later. EDIT

Location * No organisation BROWSE

Parent organisation or folder

CREATE CANCEL

3. Choose the **Navigation Tab**, then **API and Setting**, and **Create Client Id** and **ClientSecret** which will require further configuration for creating a connection between AWS Glue and GoogleAds. For more information, see API credentials.


Welcome to Google Search Con
Credentials – APIs and services
console.cloud.google.com/apis/credentials?project=myproject13jan
Google Cloud
MyProject13Jan
Search Products, resources, docs (/)
APIs and services
Credentials
+ CREATE CREDENTIALS
DELETE
Enabled APIs and services
Library
Credentials
OAuth consent screen
Domain verification
Page usage agreements
SHOW KEY
Manage service accounts
Type here to search
27°C
12:55 PM
6/20/2022

Create credentials to access your enabled APIs. Learn more

API keys

☐	Name	Creation date ↓	Restrictions	Actions
☐	API key 1	28 Jan 2022	None	SHOW KEY

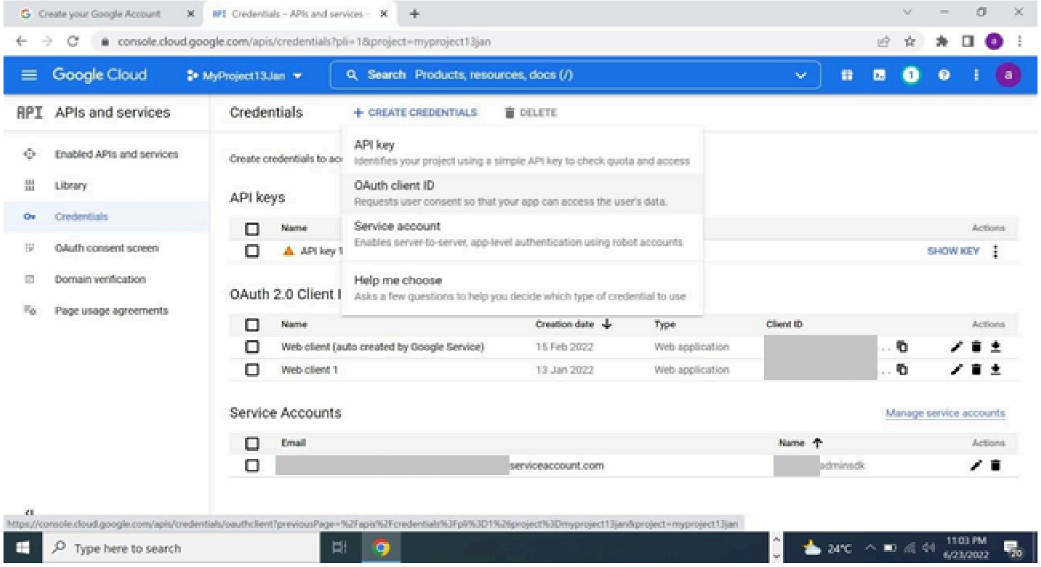
OAuth 2.0 Client IDs

☐	Name	Creation date ↓	Type	Client ID	Actions
☐	Web client (auto created by Google Service)	15 Feb 2022	Web application	[redacted] ..	
☐	Web client 1	13 Jan 2022	Web application	[redacted] ..	

Service Accounts

☐	Email	Name ↑	Actions
☐	[redacted] serviceaccount.com	[redacted] adminsdk	

4. Choose **CREATE CREDENTIALS** and choose **OAuth client ID**.


Create your Google Account
Credentials – APIs and services
console.cloud.google.com/apis/credentials?pli=1&project=myproject13jan
Google Cloud
MyProject13Jan
Search Products, resources, docs (/)
APIs and services
Credentials
+ CREATE CREDENTIALS
DELETE
Enabled APIs and services
Library
Credentials
OAuth consent screen
Domain verification
Page usage agreements
SHOW KEY
Manage service accounts
https://console.cloud.google.com/apis/credentials/oauthclient?previousPage=%2Fapis%2Fcredentials%3Fpli%3D1%26project%3Dmyproject13jan&project=myproject13jan
Type here to search
24°C
11:03 PM
6/23/2022

API key
Identifies your project using a simple API key to check quota and access

OAuth client ID
Requests user consent so that your app can access the user's data.

Service account
Enables server-to-server, app-level authentication using robot accounts

Help me choose
Asks a few questions to help you decide which type of credential to use

Create credentials to ac

API keys

☐	Name	Actions
☐	API key 1	SHOW KEY

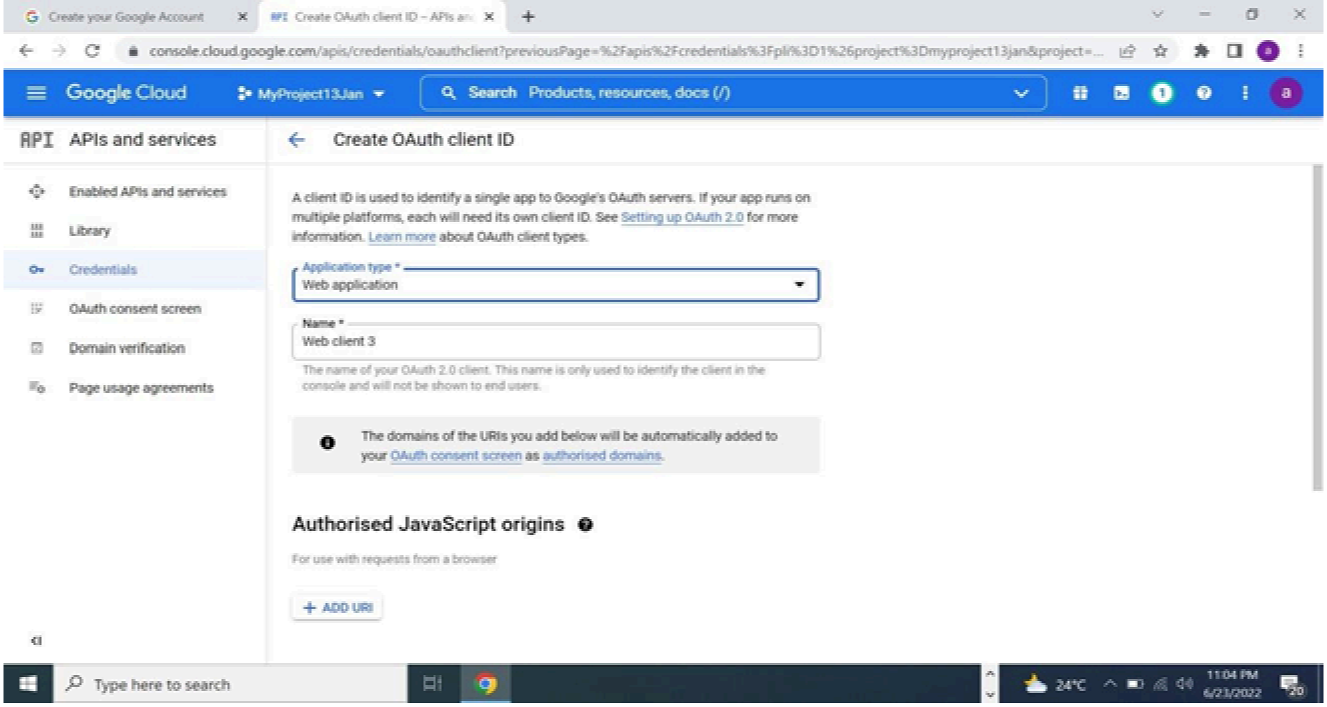
OAuth 2.0 Client I

☐	Name	Creation date ↓	Type	Client ID	Actions
☐	Web client (auto created by Google Service)	15 Feb 2022	Web application	[redacted] ..	
☐	Web client 1	13 Jan 2022	Web application	[redacted] ..	

Service Accounts

☐	Email	Name ↑	Actions
☐	[redacted] serviceaccount.com	[redacted] adminsdk	

5. Select the **Application type** as **Web application**.


Create your Google Account
Create OAuth client ID – APIs an
console.cloud.google.com/apis/credentials/oauthclient?previousPage=%2Fapis%2Fcredentials%3Fpli%3D1%26project%3Dmyproject13jan&project=...
Google Cloud
MyProject13Jan
Search Products, resources, docs (/)
APIs and services
Enabled APIs and services
Library
Credentials
OAuth consent screen
Domain verification
Page usage agreements
Create OAuth client ID
+ ADD URI
Type here to search
24°C
11:04 PM
6/23/2022

A client ID is used to identify a single app to Google's OAuth servers. If your app runs on multiple platforms, each will need its own client ID. See Setting up OAuth 2.0 for more information. Learn more about OAuth client types.

Application type * Web application

Name * Web client 3

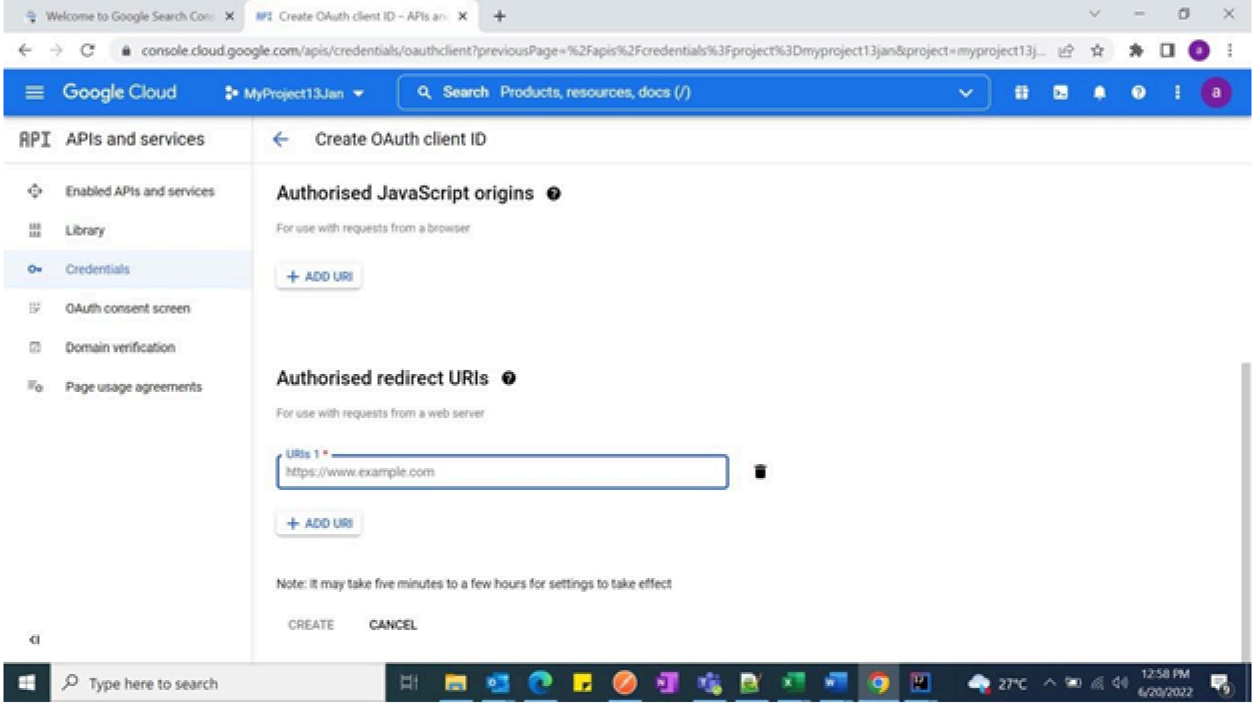
The name of your OAuth 2.0 client. This name is only used to identify the client in the console and will not be shown to end users.

The domains of the URIs you add below will be automatically added to your OAuth consent screen as authorised domains.

Authorised JavaScript origins

For use with requests from a browser

6. Under **Authorised redirect URIs**, add the OAuth Redirect URIs and choose **Create**. You can add multiple redirect URIs if required.


Welcome to Google Search Con
Create OAuth client ID – APIs an
console.cloud.google.com/apis/credentials/oauthclient?previousPage=%2Fapis%2Fcredentials%3Fproject%3Dmyproject13jan&project=myproject13j...
Google Cloud
MyProject13Jan
Search Products, resources, docs (/)
APIs and services
Enabled APIs and services
Library
Credentials
OAuth consent screen
Domain verification
Page usage agreements
Create OAuth client ID
+ ADD URI
+ ADD URI
CREATE
CANCEL
Type here to search
27°C
12:58 PM
6/20/2022

Authorised JavaScript origins

For use with requests from a browser

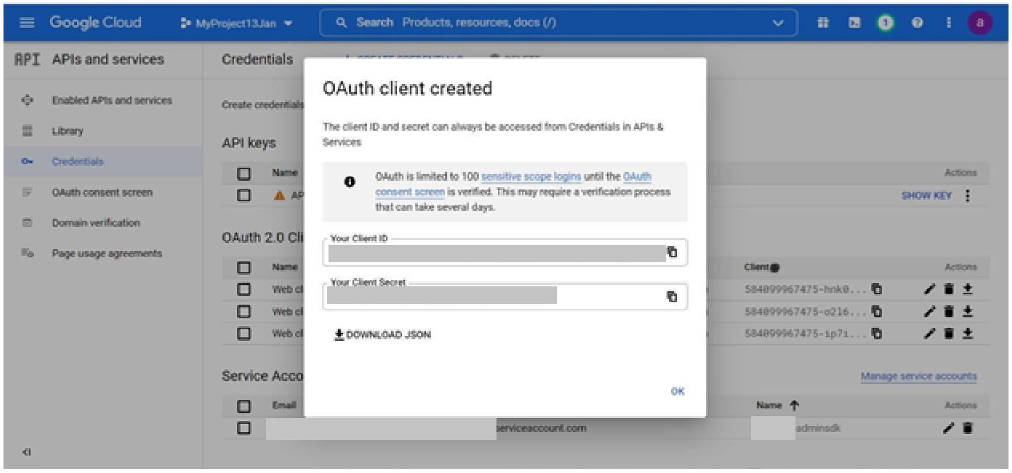
Authorised redirect URIs

For use with requests from a web server

URIs 1 * https://www.example.com

Note: It may take five minutes to a few hours for settings to take effect

7. Your **Client Id** and **Client Secret** will be generated when creating a connection between AWS Glue and Google Ads.


Google Cloud
MyProject13Jan
Search Products, resources, docs (/)
APIs and services
Enabled APIs and services
Library
Credentials
OAuth consent screen
Domain verification
Page usage agreements
Credentials
Create credentials
API keys
OAuth 2.0 Cli
Service Acco
Manage service accounts
SHOW KEY
DOWNLOAD JSON
OK

OAuth client created

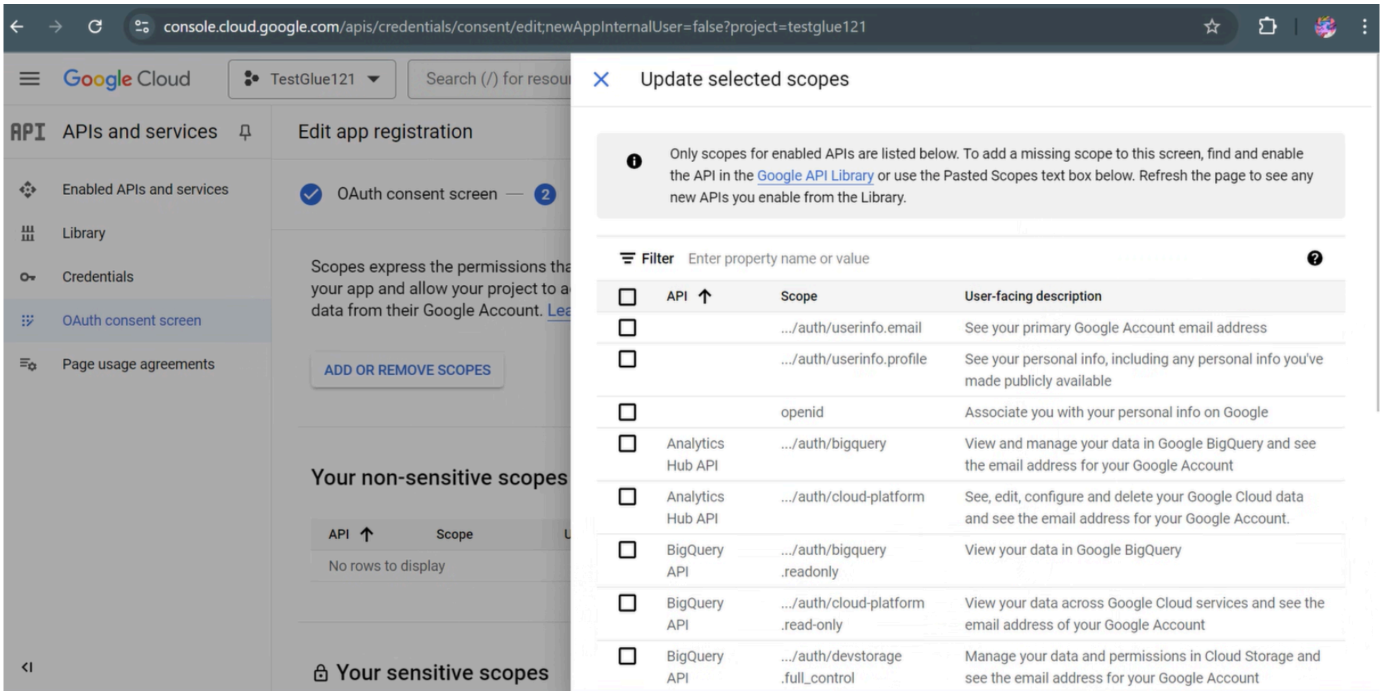
The client ID and secret can always be accessed from Credentials in APIs & Services

OAuth is limited to 100 sensitive scope logins until the OAuth consent screen is verified. This may require a verification process that can take several days.

Your Client ID: [redacted]

Your Client Secret: [redacted]

8. Add the scopes according to your application need based, choose **OAuth consent screen** and provide the required information and add the scopes based on requirements.


console.cloud.google.com/apis/credentials/consent/edit;newAppInternalUser=false?project=testglue121
Google Cloud
TestGlue121
Search (/) for resou
APIs and services
Enabled APIs and services
Library
Credentials
OAuth consent screen
Page usage agreements
Edit app registration
OAuth consent screen — 2
ADD OR REMOVE SCOPES
Your non-sensitive scopes
No rows to display
Your sensitive scopes
Filter
Enter property name or value

Scopes express the permissions tha your app and allow your project to a data from their Google Account. Lea

Update selected scopes

Only scopes for enabled APIs are listed below. To add a missing scope to this screen, find and enable the API in the Google API Library or use the Pasted Scopes text box below. Refresh the page to see any new APIs you enable from the Library.

☐	API ↑	Scope	User-facing description
☐		.../auth/userinfo.email	See your primary Google Account email address
☐		.../auth/userinfo.profile	See your personal info, including any personal info you've made publicly available
☐		openid	Associate you with your personal info on Google
☐	Analytics Hub API	.../auth/bigquery	View and manage your data in Google BigQuery and see the email address for your Google Account
☐	Analytics Hub API	.../auth/cloud-platform	See, edit, configure and delete your Google Cloud data and see the email address for your Google Account.
☐	BigQuery API	.../auth/bigquery .readonly	View your data in Google BigQuery
☐	BigQuery API	.../auth/cloud-platform .read-only	View your data across Google Cloud services and see the email address of your Google Account
☐	BigQuery API	.../auth/devstorage .full_control	Manage your data and permissions in Cloud Storage and see the email address for your Google Account

Limitations

The following are limitations for the Google Ads connector:

- `MANAGER_ID` is an optional input when creating a connection. But when you want to access the customers underlying any particular manager, then `MANAGER_ID` is a mandatory input. The table below explains the access limitations based on whether `MANAGER_ID` is included or not in a connection.

MANAGER_ID provided when creating connection?	Accessible customers
Yes	The customers listed under the provided MANAGER_ID along with the MANAGER_ID.
No	All customers will be listed, but customers underlying any manager will not be accessible.

- When a manager account is chosen as the object, only `Account` will appear as a sub-object. In the Google Ads connector, entities such as campaigns, ads, etc., are retrieved based on individual client accounts, not the manager account.
- You cannot retrieve metrics for the manager account. You can retrieve metrics for individual client accounts instead.
- Each account can have up to 10,000 campaigns, including both active and paused campaigns. For more information, see Campaign per account.
- When creating a report, if you choose certain metrics to display, any rows whose selected metrics are all zero won't be returned. For more information, see Zero Metrics .
- With the following fields, the Full Mapping flow will not work for Account, Ad Group and Ad Group Ad entities, specifically for conversionAction, conversionActionCategory, conversionActionName. For more information, see Segment and Metrics .
- A date range filter is mandatory when the `segments.date` field is selected.

Connecting to Google Analytics 4

Google Analytics 4 is an analytics service that tracks and reports metrics about visitor interactions with your apps and websites. These metrics include page views, active users, and events. If you are a Google Analytics 4 user, you can connect AWS Glue to your Google Analytics 4 account. You can use Google Analytics 4 as a data source in your ETL jobs. Run these jobs to transfer data from Google Analytics 4 to AWS services or other supported applications.

Topics

- AWS Glue support for Google Analytics 4
- Policies containing the API operations for creating and using connections

- Configuring Google Analytics 4
- Configuring Google Analytics 4 connections
- Reading from Google Analytics 4 entities
- Google Analytics 4 connection options
- Creating a Google Analytics 4 account
- Steps to create a client app and OAuth 2.0 credentials
- Limitations and considerations

AWS Glue support for Google Analytics 4

AWS Glue supports Google Analytics 4 as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Google Analytics 4.

Supported as a target?

No.

Supported Google Analytics 4 API versions

v1 Beta.

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
                "glue:ListEntities",
                "glue:DescribeEntity"
            ],
            "Resource": "*"
        }
    ]
}
```

You can also use the following managed IAM policies to allow access:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Google Analytics 4

Before you can use AWS Glue to transfer from Google Analytics 4, you must meet these requirements:

Minimum requirements

- You have a Google Analytics account with one or more data streams that collect the data that you want to transfer.
- You have a Google Cloud Platform account and a Google Cloud project.
- In your Google Cloud project, you've enabled the following APIs:
  - Google Analytics API
  - Google Analytics Admin API
  - Google Analytics Data API
- In your Google Cloud project, you've configured an OAuth consent screen for external users. For information about the OAuth consent screen, see Setting up your OAuth consent screen in the Google Cloud Platform Console Help.

- In your Google Cloud project, you've configured an OAuth 2.0 client ID. For more information, see Setting up OAuth 2.0 .

If you meet these requirements, you’re ready to connect AWS Glue to your Google Analytics 4 account.

Configuring Google Analytics 4 connections

To configure a Google Sheet connection:

1. In AWS Secrets Manager, create a secret with the following details. It is required to create a secret for each connection in AWS Glue.
   a. For AuthorizationCode grant type:
      - For customer managed connected app – Secret should contain the connected app Consumer Secret with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as key.
2. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Connection type**, select Google Analytics 4.
   b. Provide the `INSTANCE_URL` of the Google Analytics 4 you want to connect to.
   c. Select the IAM role which AWS Glue can assume and has permissions for following actions:

      JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

   d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.
   e. Select the network options if you want to use your network.
3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

`AUTHORIZATION_CODE` grant type.

This grant type is considered “three-legged” OAuth as it relies on redirecting users to the third party authorization server to authenticate the user. It is used when creating connections via the AWS Glue Console. The AWS Glue Console will redirect the user to Google Analytics 4 where the user must login and allow AWS Glue the requested permissions to access their Google Analytics 4 instance.

Users may still opt to create their own connected app in Google Analytics 4 and provide their own client id and client secret when creating connections through the AWS Glue Console. In this scenario, they will still be redirected to Google Analytics 4 to login and authorize AWS Glue to access their resources.

This grant type results in a refresh token and access token. The access token is short lived, and may be refreshed automatically without user interaction using the refresh token.

For more information, see Using Auth 2.0 to Access Google APIs .

Reading from Google Analytics 4 entities

Prerequisites

- A Google Analytics 4 object you would like to read from. Refer the supported entities table below to check the available entities.

Supported entities

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Real-Time Report	Yes	Yes	Yes	Yes	No

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Core Report	Yes	Yes	Yes	Yes	Yes

Example

```
googleAnalytics4_read = glueContext.create_dynamic_frame.from_options(
    connection_type="GoogleAnalytics4",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "entityName",
        "API_VERSION": "v1beta"
    }
```

Google Analytics 4 entity and field details

Entity	Field	Data Type	Supported Operators
Core Report	Dynamic Fields		
Core Report	Dimension Fields	String	LIKE, =
Core Report	Dimension Fields	Date	LIKE, =
Core Report	Metric Fields	String	>, <, >=, <=, = BETWEEN
Core Report	Custom Dimension and Custom Metric Fields	String	NA
Real-Time Report	appVersion	String	LIKE, =
Real-Time Report	audienceId	String	LIKE, =
Real-Time Report	audienceName	String	LIKE, =

Entity	Field	Data Type	Supported Operators
Real-Time Report	city	String	LIKE, =
Real-Time Report	cityId	String	LIKE, =
Real-Time Report	country	String	LIKE, =
Real-Time Report	countryId	String	LIKE, =
Real-Time Report	deviceCategory	String	LIKE, =
Real-Time Report	eventName	String	LIKE, =
Real-Time Report	minutesAgo	String	LIKE, =
Real-Time Report	platform	String	LIKE, =
Real-Time Report	streamId	String	LIKE, =
Real-Time Report	streamName	String	LIKE, =
Real-Time Report	unifiedScreenName	String	LIKE, =
Real-Time Report	activeUsers	String	>, <, >=, <=, = BETWEEN
Real-Time Report	conversions	String	>, <, >=, <=, = BETWEEN
Real-Time Report	eventCount	String	>, <, >=, <=, = BETWEEN
Real-Time Report	screenPageViews	String	>, <, >=, <=, = BETWEEN

Partitioning queries

1. **Filter-based partition**

Additional spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, `NUM_PARTITIONS` can be provided if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition query.
- `LOWER_BOUND`: an inclusive lower bound value of the chosen partition field.

  For date, we accept the Spark date format used in Spark SQL queries. Example of valid values: `"2024-02-06"`.
- `UPPER_BOUND`: an exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: number of partitions.

Example

```
googleAnalytics4_read = glueContext.create_dynamic_frame.from_options(
    connection_type="GoogleAnalytics4",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "entityName",
        "API_VERSION": "v1beta",
        "PARTITION_FIELD": "date"
        "LOWER_BOUND": "2022-01-01"
        "UPPER_BOUND": "2024-01-02"
        "NUM_PARTITIONS": "10"
    }
```

2. **Record-based partition**

   Additional spark options `NUM_PARTITIONS` can be provided if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by spark tasks concurrently.

   - `NUM_PARTITIONS`: number of partitions.

   Example

```
googleAnalytics4_read = glueContext.create_dynamic_frame.from_options(
    connection_type="GoogleAnalytics4",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "entityName",
        "API_VERSION": "v1beta",
        "NUM_PARTITIONS": "10"
    }
```

Google Analytics 4 connection options

The following are connection options for Google Analytics 4:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your Object in Google Analytics 4.
- `API_VERSION`(String) - (Required) Used for Read. Google Analytics 4 Rest API version you want to use.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.
- `INSTANCE_URL`(Integer) - Used for Read. (Optional)

Creating a Google Analytics 4 account

Follow the steps to create a Google Analytics 4 account: https://support.google.com/analytics/answer/9304153?hl=en

Steps to create a client app and OAuth 2.0 credentials

For more information, see Google Analytics4 API documentation.

1. Create and set up your account by logging in to your [Google Analytics Account](https://analytics.google.com) with your credentials. Then navigate to **Admin** > **Create Account**.
2. Create property for the account you have created by choosing **Create Property**. Set up the property with required details. Once all the details provided corresponding property id will be generated.
3. Add Data Stream for the created property by choosing **Data Streams** > **Add Stream** > **Web** from the drop-down. Provide the website details such as URL and other required fields. After providing all details, the corresponding **stream id** and **measurement id** will be generated.
4. Set up Google Analytics in your website by copying the measurement id and add to your website's configuration.
5. Create Report from Google Analytics by navigating to **Reports** and generating the required report.
6. Authorize your app by navigating to console.cloud.google.com and search for Google Analytics Data API, then enable the API.
   1. Navigate to the API and Services page and choose **Credentials** > **setup OAuth 2.0 Client IDs**.
   2. Provide redirect URL by adding the AWS Glue Redirect URL.
7. Copy the client id and client secret which will require further for creating connection.

Limitations and considerations

The following are limitations for the Google Analytics 4 connector:

- For the Core Report entity, only 9 dimension fields and 10 metric fields are allowed to send in a request. If the allowed number of fields is exceeded then request will fail and connector will throw an error message.
- For the Realtime Report entity, only 4 dimension fields are allowed to send in a request. If the allowed number of fields is exceeded then request will fail and connector will throw an error message.
- Google Analytics 4 is a beta version free tool, so there will be regular update on new feature, entities enhancement, adding new fields and deprecating existing fields.
- Core Report fields are populated dynamically, so there will be addition, depreciation and renaming of fields and imposing new limits on fields can be done anytime.

- The default start date is 30 days and the end date is yesterday (one day before the current date), and these dates will get overridden in filter expression code if user has set the value OR if the flow is incremental.
- As per the documentation, Real-Time report entity returns 10,000 records if limit is not pass in the request, otherwise the API returns a maximum of 250,000 rows per request, no matter how many you ask for. For more information see [Method: properties.runRealtimeReport] in the Google Analytics documentation.
- Real-Time Report entity does not support Record Based Partition as it does not support pagination. Also, it does not support Field Based Partition as none of the fields fulfill the criteria defined.
- Due to the limitation on number of fields that can be passed in a request. We are setting default dimension and metric fields within the designated limits. If "select all" is chosen, only the data from those predetermined fields will be retrieved.
  - Core Report
    - As per limitation from SAAS - requests are allowed up to 9 dimensions and up to 10 metrics only (that is, a request can contain a maximum of 19 fields(metrics + dimension).
    - As per the implementation - If user utilizes SELECT_ALL or selected fields more than 25, then default fields will be pass in the request.
    - The following fields are considered as default fields for Core Report - "country", "city", "eventName", "cityId", "browser", "date", "currencyCode", "deviceCategory", "transactionId", active1DayUsers", "active28DayUsers", "active7DayUsers", "activeUsers", "averagePurchaseRevenue", "averageRevenuePerUser", "averageSessionDuration", "engagedSessions", "eventCount", "engagementRate".
  - Real-Time Report
    - As per limitation from SAAS requests are allowed up to 4 dimensions.
    - If user pass SELECT_ALL or selected fields more than 15, then default fields will be pass in the request.
    - The following fields are considered as default fields for RealTime Report - "country", "deviceCategory", "city", "cityId", "activeUsers", "conversions", "eventCount", "screenPageViews".
- In Core-Report entity, if partition on date field and filter on startDate is present simultaneously. In that case dateRange value gets overridden with the startDate filter value, But, since partition must always be the priority, hence discarding startDate filter if partition on date field is already present.

- As now cohortSpecs is also a part of core-report request body we enhanced the current core-report entity to include support for the cohortSpec attribute. In cohortSpecs request body, nearly all fields require user input. To address this, we have set default values for those attributes/fields and provided provision for user to override these values if needed.

FieldName	Default values	Sample query to pass in filterPredicate options to override default values
startDate	30 days ago from the current date	"startDate between "2023-05-09" and "2023-05-10"
endDate	1 day ago from current date	"startDate between "2023-05-09" and "2023-05-10"
startOffset	0	startOffset=2
endOffset	1	endOffset=10
granularity	DAILY	granularity="WEEKLY"

- You can also pass all these filters together at once or with other filters.
  - Example 1 - filterPredicate: startDate between "2023-05-09" and "2023-05-10" AND startOffset=1 AND endOffset=2 AND granularity="WEEKLY"
  - Example 2 - filterPredicate: city=“xyz” AND startOffset=1 AND endOffset=2 AND granularity="WEEKLY"
- In cohort request:
  - If ‘cohortNthMonth’ is passed in the request, then internally granularity value will be set as “MONTHLY”
  - Similarly, if ‘cohortNthWeek’ is passed, then granularity value will be set as “WEEKLY”
  - And, for ‘cohortNthDay’ granularity value will be set as “DAILY”. For more information, see:
    - https://developers.google.com/analytics/devguides/reporting/data/v1/advanced
    - https://developers.google.com/analytics/devguides/reporting/data/v1/rest/v1beta/CohortSpec

- Provision is given for the user to override dateRange and granularity default value. Refer to the above table.

Connecting to Google BigQuery in AWS Glue Studio

Note

You can use AWS Glue for Spark to read from and write to tables in Google BigQuery in AWS Glue 4.0 and later versions. To configure Google BigQuery with AWS Glue jobs programmatically, see
BigQuery connections.

AWS Glue Studio provides a visual interface to connect to BigQuery, author data integration jobs, and run them on the AWS Glue Studio serverless Spark runtime.

When creating a connection to Google BigQuery in AWS Glue Studio, a unified connection is created. For more information, see Considerations.

Instead of creating a secret with the credentials in a specific format, `{"credentials": "base64 encoded JSON"}`, now with unified connection to Google BigQuery, you can create a secret which directly includes the JSON from Goolge BigQuery: `{"type": "service-account", ...}`.

Topics

- Creating a BigQuery connection
- Creating a BigQuery source node
- Creating a BigQuery target node
- Advanced options

Creating a BigQuery connection

To connect to Google BigQuery from AWS Glue, you will need to create and store your Google Cloud Platform credentials in a AWS Secrets Manager secret, then associate that secret with a Google BigQuery AWS Glue connection.

To configure a connection to BigQuery:

1. In Google Cloud Platform, create and identify relevant resources:
   - Create or identify a GCP project containing BigQuery tables you would like to connect to.
   - Enable the BigQuery API. For more information, see [Use the BigQuery Storage Read API to read table data]().
2. In Google Cloud Platform, create and export service account credentials:

   You can use the BigQuery credentials wizard to expedite this step: [Create credentials]().

   To create a service account in GCP, follow the tutorial available in [Create service accounts]().

   - When selecting **project**, select the project containing your BigQuery table.
   - When selecting GCP IAM roles for your service account, add or create a role that would grant appropriate permissions to run BigQuery jobs to read, write or create BigQuery tables.

   To create credentials for your service account, follow the tutorial available in [Create a service account key]().

   - When selecting key type, select **JSON**.

   You should now have downloaded a JSON file with credentials for your service account. It should look similar to the following:

```
{
  "type": "service_account",
  "project_id": "*****",
  "private_key_id": "*****",
  "private_key": "*****",
  "client_email": "*****",
  "client_id": "*****",
  "auth_uri": "https://accounts.google.com/o/oauth2/auth",
  "token_uri": "https://oauth2.googleapis.com/token",
  "auth_provider_x509_cert_url": "https://www.googleapis.com/oauth2/v1/certs",
  "client_x509_cert_url": "*****",
  "universe_domain": "googleapis.com"
}
```

3. In AWS Secrets Manager, create a secret using your downloaded credentials file. You can choose the **Plaintext** tab and paste the JSON formatted file content. To create a secret in Secrets Manager, follow the tutorial available in [Create an AWS Secrets Manager secret](https://docs.aws.amazon.com/secretsmanager/latest/userguide/create_secret.html) in the AWS Secrets Manager documentation. After creating the secret, keep the Secret name, *`secretName`* for the next step.
4. In the AWS Glue Data Catalog, create a connection by following the steps in https://docs.aws.amazon.com/glue/latest/dg/console-connections.html. After creating the connection, keep the connection name, *`connectionName`*, for the next step.

    - When selecting a **Connection type**, select Google BigQuery.
    - When selecting an **AWS Secret**, provide *`secretName`*.
5. Grant the IAM role associated with your AWS Glue job permission to read *`secretName`*.
6. In your AWS Glue job configuration, provide *`connectionName`* as an **Additional network connection**.

Creating a BigQuery source node

Prerequisites needed

- A BigQuery type AWS Glue Data Catalog connection
- An AWS Secrets Manager secret for your Google BigQuery credentials, used by the connection.
- Appropriate permissions on your job to read the secret used by the connection.
- The name and dataset of the table and corresponding Google Cloud project you would like to read.

Adding a BigQuery data source

To add a Data source – BigQuery node:

1. Choose the connection for your BigQuery data source. Since you have created it, it should be available in the dropdown. If you need to create a connection, choose **Create BigQuery connection**. For more information, see [Overview of using connectors and connections](https://docs.aws.amazon.com/glue/latest/dg/connectors-chapter.html).

    Once you have chosen a connection, you can view the connection properties by clicking **View properties**.
2. Identify what BigQuery data you would like to read, then choose a **BigQuery Source** option

  - Choose a single table – allows you to pull all data from a table.
  - Enter a custom query – allows you to customize which data is retrieved by providing a query.

3. Describe the data you would like to read

   (Required) set **Parent Project** to the project containing your table, or a billing parent project, if relevant.

   If you chose a single table, set **Table** to the name of a Google BigQuery table in the following format: `[dataset].[table]`

   If you chose a query, provide it to **Query**. In your query, refer to tables with their fully qualified table name, in the format: `[project].[dataset].[tableName]`.

4. Provide BigQuery properties

   If you chose a single table, you do not need to provide additional properties.

   If you chose a query, you must provide the following **Custom Google BigQuery properties**:

   - Set `viewsEnabled` to true.
   - Set `materializationDataset` to a dataset. The GCP principal authenticated by the credentials provided through the AWS Glue connection must be able to create tables in this dataset.

Creating a BigQuery target node

Prerequisites needed

- A BigQuery type AWS Glue Data Catalog connection
- An AWS Secrets Manager secret for your Google BigQuery credentials, used by the connection.
- Appropriate permissions on your job to read the secret used by the connection.
- The name and dataset of the table and corresponding Google Cloud project you would like to write to.

Adding a BigQuery data target

To add a Data target – BigQuery node:

1. Choose the connection for your BigQuery data target. Since you have created it, it should be available in the dropdown. If you need to create a connection, choose **Create BigQuery connection**. For more information, see Overview of using connectors and connections .

   Once you have chosen a connection, you can view the connection properties by clicking **View properties**.
2. Identify what BigQuery table you would like to write to, then choose a **Write method**.

   - Direct – writes to BigQuery directly using the BigQuery Storage Write API.
   - Indirect – writes to Google Cloud Storage, then copies to BigQuery.

   If you would like to write indirectly, provide a destination GCS location with **Temporary GCS bucket**. You will need to provide additional configuration in your AWS Glue connection. For more information, see Using indirect write with Google BigQuery.
3. Describe the data you would like to read

   (Required) set **Parent Project** to the project containing your table, or a billing parent project, if relevant.

   If you chose a single table, set **Table** to the name of a Google BigQuery table in the following format: `[dataset].[table]`

Advanced options

You can provide advanced options when creating a BigQuery node. These options are the same as those available when programming AWS Glue for Spark scripts.

See BigQuery connection option reference in the AWS Glue developer guide.

Connecting to Google Search Console

Google Search Console is a free platform available to website owners for monitoring how Google views the site and optimizing its organic presence. This includes viewing referring domains, mobile site performance, rich search results, and the highest-traffic queries and pages. If you are a Google Search Console user, you can connect AWS Glue to your Google Search Console account. You can use Google Search Console as a data source in your ETL jobs. Run these jobs to transfer data from Google Search Console to AWS services or other supported applications.

Topics

- AWS Glue support for Google Search Console
- Policies containing the API operations for creating and using connections
- Configuring Google Search Console
- Configuring Google Search Console connections
- Reading from Google Search Console entities
- Google Search Console connection options
- Google Search Console limitations

AWS Glue support for Google Search Console

AWS Glue supports Google Search Console as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Google Search Console.

Supported as a target?

No.

Supported Google Search Console API versions

The following Google Search Console API versions are supported:

- v3

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Google Search Console

Before you can use AWS Glue to transfer data from Google Search Console, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- You have a Google Search Console account.
- You have a Google Cloud Platform account and a Google Cloud project.
- In your Google Cloud project, you've enabled the Google Search Console API.

- In your Google Cloud project, you've configured an OAuth consent screen for external users. For more information, see Setting up your OAuth consent screen in the Google Cloud Platform Console Help.
- In your Google Cloud project, you've configured an OAuth 2.0 client ID. See Setting up OAuth 2.0 for the client credentials that AWS Glue uses to access your data securely when it makes authenticated calls to your account.

If you meet these requirements, you’re ready to connect AWS Glue to your Google Search Console account. For typical connections, you don't need do anything else in Google Search Console.

Configuring Google Search Console connections

Google Search Console supports the AUTHORIZATION_CODE grant type for OAuth2. The grant type determines how AWS Glue communicates with Google Search Console to request access to your data.

- This grant type is considered "three-legged" OAuth as it relies on redirecting users to a third-party authorization server to authenticate the user. It is used when creating connections via the AWS Glue console.
- Users may still opt to create their own connected app in Google Search Console and provide their own client ID and client secret when creating connections through the AWS Glue console. In this scenario, they will still be redirected to Google Search Console to login and authorize AWS Glue to access their resources.
- This grant type results in a refresh token and access token. The access token is short lived, and may be refreshed automatically without user interaction using the refresh token.
- For public Google Search Console documentation on creating a connected app for Authorization Code OAuth flow, see Using OAuth 2.0 to Access Google APIs.

To configure a Google Search Console connection:

1. In AWS Secrets Manager, create a secret with the following details:
   a. For the customer managed connected app, the Secret should contain the connected app Consumer Secret with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as key.
   b. Note: you must create a secret for your connections in AWS Glue.

1. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Connection type**, select Google Search Console.
   b. Select the AWS IAM role which AWS Glue can assume and has permissions for following actions:

      JSON

      ```
      {
        "Version": "2012-10-17",
        "Statement": [
          {
            "Effect": "Allow",
            "Action": [
              "secretsmanager:DescribeSecret",
              "secretsmanager:GetSecretValue",
              "secretsmanager:PutSecretValue",
              "ec2:CreateNetworkInterface",
              "ec2:DescribeNetworkInterfaces",
              "ec2:DeleteNetworkInterface"
            ],
            "Resource": "*"
          }
        ]
      }
      ```
   c. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.
   d. Select the network options if you want to use your network.
2. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Google Search Console entities

Prerequisite

A Google Search Console object you would like to read from. You will need the object name.

Supported entities for source:

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Search Analytics	Yes	Yes	No	Yes	No
Sites	No	No	No	Yes	No
Sitemaps	No	No	No	Yes	No

Example:

```
googleSearchConsole_read = glueContext.create_dynamic_frame.from_options(
    connection_type="googlesearchconsole",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "entityName",
        "API_VERSION": "v3"
    }
```

Google Search Console entity and field details:

Google Search Console provides endpoints to fetch metadata dynamically for supported entities. Accordingly, operator support is captured at the datatype level.

Entity	Field	Data type	Supported operators	Note
Search Analytics	keys	List	N/A	
	clicks	Double	N/A	
	impressions	Double	N/A	
	ctr	BigDecimal	N/A	For BigDecimal datatype, the value '0' is formatted as '0E-18'

Entity	Field	Data type	Supported operators	Note
	position	Double	N/A	
	start_end_date	Date	BETWEEN	Default value for `start_end_date` is between <30 days ago from the current date> AND <yesterday: that is, 1 day ago from the current date> Note: expects you to pass a UTC date value. Example: start_end_date between '2022-01-01' AND '2024-09-09'
	country	String	EQUAL_TO, NOT_EQUAL_TO, CONTAINS	Valid values are "IND", "CAN", etc.
	type	String	EQUAL_TO, NOT_EQUAL_TO	Valid values are "discover", "googleNews", "news", "image", "video", "web"

Entity	Field	Data type	Supported operators	Note
	searchAppearance	String	EQUAL_TO, NOT_EQUAL_TO, CONTAINS	See Search Appearance for a list of valid values.
	device	String	EQUAL_TO, NOT_EQUAL_TO, CONTAINS	Valid values are "DESKTOP", "MOBILE", "TABLET"
	dimensions	String	EQUAL_TO	Valid values are "country", "device"
	page	String	EQUAL_TO, NOT_EQUAL_TO, CONTAINS	
	query	String	EQUAL_TO, NOT_EQUAL_TO, CONTAINS	
	dataState	String	EQUAL_TO	Valid values are "all" and "final"
Sites	siteUrl	String	N/A	
	permissionLevel	String	N/A	
Sitemaps	path	String	N/A	
	type	String	N/A	
	lastSubmitted	DateTime	N/A	
	isPending	Boolean	N/A	

Entity	Field	Data type	Supported operators	Note
	isSitemapsIndex	Boolean	N/A	
	lastDownloaded	DateTime	N/A	
	warnings	Long	N/A	
	errors	Long	N/A	
	contents	List	N/A	

> **Note**
>
> For an updated list of valid values for filters, see the Google Search Console API docs. The field `start_end_date` is a combination of `start_date` and `end_date`.

Partitioning queries

Filter-based partitioning and record-based partitioning are not supported.

Google Search Console connection options

The following are connection options for Google Search Console:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your object in Google Search Console.
- `API_VERSION`(String) - (Required) Used for Read. Google Search Console Rest API version you want to use.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: "start_end_date between <30 days ago from current date> AND <yesterday: that is, 1 day ago from the current date>". Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: "start_end_date between <30 days ago from current date> AND <yesterday: that is, 1 day ago from the current date>" Used for Read. Full Spark SQL query.

- `INSTANCE_URL`(String) - Used for Read. A valid Google Search Console instance URL.

Google Search Console limitations

The following are limitations or notes for Google Search Console:

- Google Search Console enforces usage limits on the API. For more information, see Usage Limits.
- When no filter is passed for the `Search Analytics` entity, the API sums up all the clicks, impressions, CTR, and other data for your entire site within the specified default date range and presents it as a single record.
- To breakdown the data into smaller segments, you need to introduce dimensions to your query. Dimensions tells the API how you want to segment your data.
  - For example, if you add `filterPredicate: dimensions="country"` you'll get one record for each country where your site received traffic during the specified period.
  - Example to pass multiple dimensions: `filterPredicate: dimensions="country" AND dimensions="device" AND dimensions="page"`. In this case you'll get one row in the response for each unique combination of these three dimensions.
- Default values are set for the `start_end_date` and `dataState` fields.

Field name	Default filter expression	Expression to override default values
start_end_date	start_end_date between <30 days ago from current date> AND <yesterday: that is, 1 day ago from the current date>	start_end_date between "2024-01-01" AND "2024-05-05"
dateState	dataState="all"	dataState="final"

Connecting to Google Sheets

Google Sheets is an online spreadsheet software that allows you to organize large amounts of data, create custom reports, automate calculations, and collaborate with others.If you're a Google Sheets user, you can connect AWS Glue to your Google Sheets account. Then, you can use Google

Sheets as a data source in your ETL jobs. Run these jobs to transfer data between Google Sheets and AWS services or other supported applications.

Topics

- AWS Glue support for Google Sheets
- Policies containing the API operations for creating and using connections
- Configuring Google Sheets
- Configuring Google Sheets connections
- Reading from Google Sheets entities
- Google Sheets connection options
- Set up Authorization code OAuth flow for Google Sheets
- Limitations for Google Sheets connector

AWS Glue support for Google Sheets

AWS Glue supports Google Sheets as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Google Sheets.

Supported as a target?

No.

Supported Google Sheets API versions

Google Sheets API v4 and Google Drive API v3

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
    "Statement": [
      {
        "Effect": "Allow",
        "Action": [
          "glue:ListConnectionTypes",
          "glue:DescribeConnectionType",
          "glue:RefreshOAuth2Tokens",
          "glue:ListEntities",
          "glue:DescribeEntity"
        ],
        "Resource": "*"
      }
    ]
  }
```

You can also use the following managed IAM policies to allow access:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Google Sheets

Before you can use AWS Glue to transfer from Google Sheets, you must meet these requirements:

Minimum requirements

- You have a Google Sheets account with Email and Password.
- Your Google Sheets account is enabled for API access. All use of the Google Sheets API is available at no additional cost.
- Your Google Sheets account allows you to install connected apps. If you lack access to this functionality, contact your Google Sheets administrator.

If you meet these requirements, you're ready to connect AWS Glue to your Google Sheets account.

Configuring Google Sheets connections

To configure a Google Sheet connection:

1. In AWS Secrets Manager, create a secret with the following details:
   a. For AuthorizationCode grant type:
      - For customer managed connected app – Secret should contain the connected app Consumer Secret with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as key.
2. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Data Source**, select Google Sheets.
   b. Provide the Google Sheets environment.
      i. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.
      ii. Select the network options if you want to use your network.
3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

   JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

AUTHORIZATION_CODE Grant Type

This grant type is considered “three-legged” OAuth as it relies on redirecting users to the third party authorization server to authenticate the user. It is used when creating connections via the AWS Glue Console. The AWS Glue Console will redirect the user to Google Sheets where the user must login and allow AWS Glue the requested permissions to access their Google Sheets instance.

Users may opt to create their own connected app in Google Sheets and provide their own client id and client secret when creating connections through the AWS Glue Console. In this scenario, they will still be redirected to Google Sheets to login and authorize AWS Glue to access their resources.

This grant type results in a refresh token and access token. The access token is short lived, and may be refreshed automatically without user interaction using the refresh token.

For more information, see public Google Sheets documentation on creating a connected app for Authorization Code OAuth flow .

Reading from Google Sheets entities

Prerequisites

- A Google SpreadSheet that you would like to read from. You will need the SpreadSheet ID and tabName of the spreadsheet.

Google Sheets Entity and Field Details:

Entity	Data Type	Supported Operators
Spreadsheet	String	N/A (filter is not supported)

Example

```
googleSheets_read = glueContext.create_dynamic_frame.from_options(
    connection_type="googlesheets",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "{SpreadSheetID}#{SheetTabName}",
        "API_VERSION": "v4"
    }
```

Partitioning queries

For Record Base Partitioning only, `NUM_PARTITIONS` can be provided as additional spark options if you want to utilize concurrency in Spark. With this parameter, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by spark tasks concurrently.

Example with `NUM_PARTITIONS`

```
googlesheets_read = glueContext.create_dynamic_frame.from_options(
    connection_type="googlesheets",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "{SpreadSheetID}#{SheetTabName}",
        "API_VERSION": "v4",
        "NUM_PARTITIONS": "10"
    }
```

Google Sheets connection options

The following are connection options for Google Sheets:

- `ENTITY_NAME`(String) - (Required) Used for Read. The `SpreadSheet ID` and `sheetTabName` in Google Sheets. Example: `{SpreadSheetID}#{SheetTabName}`.
- `API_VERSION`(String) - (Required) Used for Read. Google Sheets Rest API version you want to use.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.

Set up Authorization code OAuth flow for Google Sheets

Prerequisites

- A Google account where you can sign in to use the Google Sheets app. In your Google account, Google Sheets contains the data that you want to transfer.

- A Google Cloud Platform account and a Google Cloud project. See Create Google Cloud Project for more details.

To set up your Google account and get OAuth 2.0 credentials:

1. Once the Google Cloud project is setup, enable the Google Sheets API and Google Drive APIs in the project. For the steps to enable them, see Enable and disable APIs in the API Console Help for Google Cloud Platform.
2. Next, configure an OAuth consent screen for external users. For more information about the OAuth consent screen, see Setting up your OAuth consent screen in the Google Cloud Platform Console Help.
3. In the OAuth consent screen, add the following scopes:

    - The Google Sheets API read-only scope
    - The Google Drive API read-only scope

    For more information about these scopes, see OAuth 2.0 Scopes for Google APIs in the Google Identity documentation.
4. Generate OAuth 2.0 client ID and secret. For the steps to create this client ID, see Setting up OAuth 2.0 in the Google Cloud Platform Console Help.

    The OAuth 2.0 client ID must have one or more authorized redirect URLs.

    Redirect URLs have the following format:

    - https://<aws-region>.console.aws.amazon.com/gluestudio/oauth
5. Note the client ID and client secret from the settings for your OAuth 2.0 client ID.

Limitations for Google Sheets connector

The following are limitations for the Google Sheets connector:

- Google Sheets connector does not support Filters. Hence, filter based partitioning cannot be supported.
- In Record Base Partitioning, there is no provision to return exact record count by SAAS. As a result, there can be scenarios where files with empty records are created.

- Since the Google Sheets connector does not support filter-based partitioning, `partitionField`, `lowerbound`, and `upperbound` are not valid connection options. If these options are provided, the AWS Glue job is expected to fail.
- It is essential to designate the first row of the sheet as the header row to avoid data processing issues.
  - If not provided, header row will be replaced with `Unnamed:1`, `Unnamed:2`, `Unnamed:3`... if the sheet contains data with the first row empty.
  - If header row is provided, empty column names will be replaced with `Unnamed:<number of column>`. For example, if header row is `['ColumnName1', 'ColumnName2', '', '', 'ColumnName5', 'ColumnName6']`, then it will become `['ColumnName1', 'ColumnName2', 'Unnamed:3', 'Unnamed:4', 'ColumnName5', 'ColumnName6']`.
- The Google Sheets connector does not support Incremental transfer.
- Google Sheets connector supports only String datatype.
- Duplicate headers in a sheet will be iteratively renamed with a numeric suffix. Header names provided by the user will have precedence while renaming the duplicate headers. For example, if the header row is ["Name", "", "Name", null, "Unnamed:6", ""], it will change to: ["Name", "Unnamed:2", "Name1", "Unnamed:4", "Unnamed:6", "Unnamed:61"].
- Google Sheets connector does not support spaces for a tabName.
- A folder name can't have the following special characters:
  - #
  - /

Connecting to HubSpot

HubSpot's CRM platform has all the tools and integrations you need for marketing, sales, content management, and customer service.

- Marketing Hub - Marketing software to help you grow traffic, convert more visitors, and run complete inbound marketing campaigns at scale.
- Sales Hub - Sales CRM software to help you get deeper insights into prospects, automate the tasks you have, and close more deals faster.
- Service Hub - Customer service software to help you connect with customers, exceed expectations, and turn them into promoters who grow your business.

- Operations Hub - Operations software that syncs your apps, cleans and creates customer data, and automates processes — so all your systems and teams work better together.

Topics

- AWS Glue support for HubSpot
- Policies containing the API operations for creating and using connections
- Configuring HubSpot
- Configuring HubSpot connections
- Reading from HubSpot entities
- HubSpot connection options
- Limitations and notes for HubSpot connector

AWS Glue support for HubSpot

AWS Glue supports HubSpot as follows:

Supported as a source?

Yes – Sync and Async. You can use AWS Glue ETL jobs to query data from HubSpot.

Supported as a target?

No.

Supported HubSpot API versions

The following HubSpot API versions are supported:

- v1
- v2
- v3
- v4

For entity support per version specific, see Supported entities for Sync source and Supported entities for Async source.

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring HubSpot

Before you can use AWS Glue to transfer data from HubSpot, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- You have a HubSpot account. For more information, see Creating a HubSpot account.
- Your HubSpot account is enabled for API access.
- You should have an app under your HubSpot developer account that provides the client credentials that AWS Glue uses to access your data securely when it makes authenticated calls to your account. For more information, see Creating a HubSpot developer app.

If you meet these requirements, you’re ready to connect AWS Glue to your HubSpot account. For typical connections, you don't have do anything else in HubSpot.

Creating a HubSpot account

To create a HubSpot account:

1. Go to the HubSpot CRM SignUp URL.
2. Enter your email address and choose **Verify email** (alternatively, you can choose to sign up with a Google, Microsoft or Apple account).
3. Check your inbox for the verification code from HubSpot.
4. Enter the 6-digit verification code and click **Next**.
5. Enter a password and click **Next**.
6. Enter your first name and last name and click **Next**, or sign up using the **Sign up with Google** link.
7. Enter your industry and click **Next**.
8. Enter your job role and click **Next**.
9. Enter your company name and click **Next**.
10. Select the size of your company (number of employees working in your company) and click **Next**.
11. Enter your company website and click **Next**.
12. Select where your data should be hosted (United States or Europe) and click **Create Account**.
13. Select the purpose of your account creation and click **Next**.
14. Choose **Connect Google Account** or choose to add contacts yourself to link your contacts with your HubSpot account.

15. Log in to your Google account if you chose the **Connect Google Account** option to link your contacts and start using your HubSpot account.

Creating a HubSpot developer app

App developer accounts are intended for creating and managing apps, integrations, and developer test accounts. They're also where you can create and manage App Marketplace listings. However, app developer accounts and their associated test accounts aren’t connected to a standard HubSpot account. They can't sync data or assets to or from another HubSpot account. To get the Client ID and Client Secret you create a developers account.

1. Go to https://developers.hubspot.com/
2. Choose **Create developer account** and scroll down.
3. You will be asked whether you want to create an App developers account, Private App account, or CMS Developer Sandbox account. Choose **Create App developers account**.
4. Since you already created an Account with HubSpot, you can choose **Continue with this user**.
5. Click **Start signup**.
6. Enter your Job Role and click **Next**.
7. Name your developer account and click **Next**, then **Skip**.
8. Choose **Create App**.
9. Once your App is created, choose **Auth**.
10. Under Auth, note the Client ID and Client Secret.
11. Add your region specific **Redirect URL** as https:*//<aws-region>*.console.aws.amazon.com/gluestudio/oauth. For example, add https://us-east-1.console.aws.amazon.com/gluestudio/oauth for the us-east-1 region.
12. Scroll down and find scopes. There are two kinds of scopes you must select under headings "CRM" and "Standard".
13. Add the following scopes:

```
content
automation
oauth
crm.objects.owners.read
forms
tickets
```

```
crm.objects.contacts.write
e-commerce
crm.schemas.custom.read
crm.objects.custom.read
sales-email-read
crm.objects.custom.write
crm.objects.companies.write
crm.lists.write
crm.objects.companies.read
crm.lists.read
crm.objects.deals.read
crm.objects.deals.write
crm.objects.contacts.read
```

14. Click **Save** and your dev account is now ready to use.
15. Scroll above to find the **Client ID**.
16. On the same page, click **Show** to get the **Client secret**.

Creating a HubSpot developer test account

Within app developer accounts, you can create developer test accounts to test apps and integrations without affecting any real HubSpot data. Developer test accounts do not mirror production accounts, but rather have access to a 90-day trial of the Enterprise versions of Marketing, Sales, Service, CMS, and Operations Hub, providing the ability to test most HubSpot tools and APIs.

1. Click **Home**.
2. Click **Create test account**.
3. Click **Create App Test Account**.
4. A new window appears. Enter the app test account name and click **Create**.

   Your app test account is now created.

> **Note**
>
> The developer account is related to development activities such as API integration, and the app test account is used to see the data which is created or pulled by the developer account.

Configuring HubSpot connections

HubSpot supports the AUTHORIZATION_CODE grant type for OAuth2.

- This grant type is considered "three-legged" OAuth as it relies on redirecting users to a third-party authorization server to authenticate the user. It is used when creating connections via the AWS Glue console. The user creating a connection needs to provide the OAuth related information like Client ID and Client Secret for their HubSpot client application. The AWS Glue console will redirect the user to HubSpot where the user must login and allow AWS Glue the requested permissions to access their HubSpot instance.
- Users may still opt to create their own connected app in HubSpot and provide their own client id and client secret when creating connections through the AWS Glue console. In this scenario, they will still be redirected to HubSpot to login and authorize AWS Glue to access their resources.
- This grant type results in a refresh token and access token. The access token is short lived, and may be refreshed automatically without user interaction using the refresh token.
- For public HubSpot documentation on creating a connected app for Authorization Code OAuth flow, see [Public apps](#).

To configure a HubSpot connection:

1. In AWS Secrets Manager, create a secret with the following details:
   a. For the customer managed connected app, the Secret should contain the connected app Consumer Secret with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as key.
   b. Note: You must create a secret for the connection in AWS Glue.
2. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Connection type**, select HubSpot.
   b. Provide the HubSpot environment.
   c. Select the AWS IAM role which AWS Glue can assume and has permissions for following actions:

      JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

   d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.

   e. Select the network options if you want to use your network.

3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.
4. In your AWS Glue job configuration, provide `connectionName` as an **Additional network connection**.

Reading from HubSpot entities

Prerequisite

A HubSpot object you would like to read from. You will need the object name such as contact or task. The following table shows the supported entities for Sync source.

Supported entities for Sync source

Entity	API version	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partioning
Campaigns	v1	No	Yes	No	Yes	No
Companies	v3	Yes	Yes	Yes	Yes	Yes
Contacts	v3	Yes	Yes	Yes	Yes	Yes

Entity	API version	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partioning
Contact Lists	v1	No	Yes	No	Yes	No
Deals	v3	Yes	Yes	Yes	Yes	Yes
CRM Pipeline (Deal Pipelines)	v1	No	No	No	Yes	No
Email Events	v1	No	Yes	No	Yes	No
Calls	v3	Yes	Yes	Yes	Yes	Yes
Notes	v3	Yes	Yes	Yes	Yes	Yes
Emails	v3	Yes	Yes	Yes	Yes	Yes
Meetings	v3	Yes	Yes	Yes	Yes	Yes
Tasks	v3	Yes	Yes	Yes	Yes	Yes
Postal Mails	v3	Yes	Yes	Yes	Yes	Yes
Custom Objects	v3	Yes	Yes	Yes	Yes	Yes
Forms	v2	No	No	No	Yes	No
Owners	v3	No	Yes	No	Yes	No
Products	v3	Yes	Yes	Yes	Yes	Yes
Tickets	v3	Yes	Yes	Yes	Yes	Yes
Workflows	v3	No	No	No	Yes	No

Entity	API version	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partioning
Associations	v4	Yes	No	No	Yes	No
Associations Labels	v4	No	No	No	Yes	No

Example:

```
hubspot_read = glueContext.create_dynamic_frame.from_options(
    connection_type="hubspot",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "contact",
        "API_VERSION": "v3"
    }
```

Supported entities for Async source

Entity	API version	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partioning
Companies	v3	Yes	No	Yes	Yes	No
Contacts	v3	Yes	No	Yes	Yes	No
Deals	v3	Yes	No	Yes	Yes	No
Calls	v3	Yes	No	Yes	Yes	No
Notes	v3	Yes	No	Yes	Yes	No
Emails	v3	Yes	No	Yes	Yes	No
Meetings	v3	Yes	No	Yes	Yes	No
Tasks	v3	Yes	No	Yes	Yes	No

Entity	API version	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partioning
Postal Mails	v3	Yes	No	Yes	Yes	No
Custom Objects	v3	Yes	No	Yes	Yes	No
Products	v3	Yes	No	Yes	Yes	No
Tickets	v3	Yes	No	Yes	Yes	No

Example:

```
hubspot_read = glueContext.create_dynamic_frame.from_options(
    connection_type="hubspot",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "contact",
        "API_VERSION": "v3",
        "TRANSFER_MODE": "ASYNC"
    }
```

HubSpot entity and field details:

HubSpot API v4:

Entity	API version	Field	Data type	Supported operators
Association Label	v4	category	String	N/A
		typeId	Integer	N/A
		label	String	N/A
Associations		from	Struct	N/A
		id	String	"="

Entity	API version	Field	Data type	Supported operators
		to	List	N/A

Note

For the `Associations` object, to fetch associations between two objects, you need to provide the 'from Id' (the ID of the first object) via a mandatory filter while creating an AWS Glue job. If you want to fetch associations for multiple from IDs in that case, you have to provide multiple IDs in the `where` clause. For example: for fetching `Associations` for contact IDs '1' and '151', you need to provide a filter as `where id=1 AND id=151`.

HubSpot API v3:

Entity	Field	Data type	Supported operators
Owner	firstName	String	N/A
	lastName	String	N/A
	createdAt	DateTime	N/A
	archived	Boolean	N/A
	teams	List	N/A
	id	String	N/A
	userId	Integer	N/A
	email	String	N/A
	updatedAt	DateTime	N/A
Workflow	name	String	N/A
	id	Integer	N/A

Entity	Field	Data type	Supported operators
	type	String	N/A
	enabled	Boolean	N/A
	insertedAt	Long	N/A
	updatedAt	Long	N/A
	contactListIds	Struct	N/A
	personaTagIds	List	N/A

For the following entities, HubSpot provides endpoints to fetch metadata dynamically, so that operator support is captured at the datatype level for each entity.

Note

`DML_STATUS` is a virtual field added on every record at runtime to determine its status (CREATED/UPDATED) in the Sync mode. The `CONTAINS/LIKE` operator is not supported in the Async mode.

Entity	Data type	Supported operators
Contact	Integer	"=, !=, <, >, >=, <="
	Long	"=, !=, <, >, >=, <="
	String	"=, !=, LIKE"
	Date	N/A
	DateTime	"between"
	Boolean	"="
	List	N/A

Entity	Data type	Supported operators
	Struct	N/A
Company	Integer	"=, !=, <, >, >=, <="
	Long	"=, !=, <, >, >=, <="
	String	"=, !=, LIKE"
	Date	N/A
	DateTime	"between"
	Boolean	"="
	List	N/A
	Struct	N/A
Deal	Integer	"=, !=, <, >, >=, <="
	Long	"=, !=, <, >, >=, <="
	String	"=, !=, LIKE"
	Date	N/A
	DateTime	"between"
	Boolean	"="
	List	N/A
	Struct	N/A
Ticket	Integer	"=, !=, <, >, >=, <="
	Long	"=, !=, <, >, >=, <="
	String	"=, !=, LIKE"

Entity	Data type	Supported operators
	Date	N/A
	DateTime	"between"
	Boolean	"="
	List	N/A
	Struct	N/A
Product	Integer	"=, !=, <, >, >=, <="
	Long	"=, !=, <, >, >=, <="
	String	"=, !=, LIKE"
	Date	N/A
	DateTime	"between"
	Boolean	"="
	List	N/A
	Struct	N/A
Custom Object	Integer	"=, !=, <, >, >=, <="
	Long	"=, !=, <, >, >=, <="
	String	"=, !=, LIKE"
	Date	N/A
	DateTime	"between"
	Boolean	"="
	List	N/A

Entity	Data type	Supported operators
	Struct	N/A
Call	Integer	"=, !=, <, >, >=, <="
	Long	"=, !=, <, >, >=, <="
	String	"=, !=, LIKE"
	Date	N/A
	DateTime	"between"
	Boolean	"="
	List	N/A
	Struct	N/A
Email	Integer	"=, !=, <, >, >=, <="
	Long	"=, !=, <, >, >=, <="
	String	"=, !=, LIKE"
	Date	N/A
	DateTime	"between"
	Boolean	"="
	List	N/A
	Struct	N/A
Meeting	Integer	"=, !=, <, >, >=, <="
	Long	"=, !=, <, >, >=, <="
	String	"=, !=, LIKE"

Entity	Data type	Supported operators
	Date	N/A
	DateTime	"between"
	Boolean	"="
	List	N/A
	Struct	N/A
Note	Integer	"=, !=, <, >, >=, <="
	Long	"=, !=, <, >, >=, <="
	String	"=, !=, LIKE"
	Date	N/A
	DateTime	"between"
	Boolean	"="
	List	N/A
	Struct	N/A
Task	Integer	"=, !=, <, >, >=, <="
	Long	"=, !=, <, >, >=, <="
	String	"=, !=, LIKE"
	Date	N/A
	DateTime	"between"
	Boolean	"="
	List	N/A

Entity	Data type	Supported operators
	Struct	N/A
Postal Mail	Integer	"=, !=, <, >, >=, <="
	Long	"=, !=, <, >, >=, <="
	String	"=, !=, LIKE"
	Date	N/A
	DateTime	"between"
	Boolean	"="
	List	N/A
	Struct	N/A

HubSpot API v2:

Entity	Field	Data type	Supported operators
Form	portalId	Integer	N/A
	guid	String	N/A
	name	String	N/A
	method	String	N/A
	cssClass	String	N/A
	redirect	String	N/A
	submitText	String	N/A
	notifyRecipients	String	N/A

Entity	Field	Data type	Supported operators
	createdAt	Long	N/A
	updatedAt	Long	N/A
	ignoreCurrentValues	Boolean	N/A
	deletable	Boolean	N/A
	inlineMessage	Boolean	N/A
	captchaEnabled	Boolean	N/A
	cloneable	Boolean	N/A
	formFieldGroups	List	N/A
	editable	Boolean	N/A
	deletedAt	Integer	N/A
	themeName	String	N/A
	parentId	Integer	N/A
	style	String	N/A
	isPublished	Boolean	N/A
	publishAt	Integer	N/A
	unpublishAt	Integer	N/A
	publishedAt	Integer	N/A
	kickbackEmailWorkflowId	String	N/A
	kickbackEmailsJson	Integer	N/A
	customUid	String	N/A

Entity	Field	Data type	Supported operators
	createMarketableContact	Boolean	N/A
	editVersion	Integer	N/A
	thankYouMessageJson	String	N/A
	themeColor	String	N/A
	alwaysCreateNewCompany	Boolean	N/A
	internalUpdatedAt	Long	N/A
	businessUnitId	Integer	N/A
	portableKey	String	N/A
	paymentSessionTemplateIds	List	N/A
	selectedExternalOptions	List	N/A

HubSpot API v1:

Entity	Field	Data type	Supported operators
Campaign	id	Integer	N/A
	appId	Integer	N/A
	appName	String	N/A
	lastUpdatedTime	Long	N/A
Contact_List	dynamic	Boolean	N/A

Entity	Field	Data type	Supported operators
	name	String	N/A
	portalId	Integer	N/A
	createdAt	Long	N/A
	listId	Integer	N/A
	updatedAt	Long	N/A
	ListType	String	N/A
	filters	List	N/A
	authorId	Integer	N/A
	metaData	Struct	N/A
	archived	Boolean	N/A
	ilsFilterBranch	String	N/A
	filterIds	List	N/A
	limitExempt	Boolean	N/A
	internal	Boolean	N/A
	readOnly	Boolean	N/A
	parentId	Integer	N/A
Email_Event	id	String	N/A
	type	String	N/A
	recipient	String	N/A
	portalId	Integer	N/A

Entity	Field	Data type	Supported operators
	appId	Integer	N/A
	appName	String	N/A
	emailCampaignId	Long	N/A
	attempt	Integer	N/A
	created	Long	N/A
	sentBy	Struct	N/A
	smtpId	String	N/A
	response	String	N/A
	subject	String	N/A
	cc	List	N/A
	bcc	List	N/A
	replyTo	List	N/A
	from	String	N/A
	dropReason	String	N/A
	dropMessage	String	N/A
	browser	Struct	N/A
	userAgent	String	N/A
	duration	Long	N/A
	location	Struct	N/A
	filteredEvent	Boolean	N/A

Entity	Field	Data type	Supported operators
	deviceType	String	N/A
	suppressedReason	String	N/A
	suppressedMessage	String	N/A
CRM_Pipeline	pipelineId	String	N/A
	createdAt	Long	N/A
	updatedAt	Long	N/A
	objectType	String	N/A
	label	String	N/A
	displayOrder	Integer	N/A
	active	Boolean	N/A
	stages	List	N/A
	objectTypeId	String	N/A
	default	Boolean	N/A

Partitioning queries

You can provide the additional Spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, and `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition the query.
- `LOWER_BOUND`: an **inclusive** lower bound value of the chosen partition field.

  For the DateTime field, we accept the value in ISO format.

  Examples of valid value:

```
“2024-01-01T10:00:00.115Z"
```

- `UPPER_BOUND`: an **exclusive** upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: the number of partitions.

The following table describes the entity partitioning field support details:

Entity name	Partitioning fields	Data type
contact	hs_object_id	Long
	createdate, lastmodifieddate	DateTime
company	hs_object_id	Long
	createdate, hs_lastmodifieddate	DateTime
deal	hs_object_id	Long
	createdate, hs_createdate, hs_lastmodifieddate	DateTime
ticket	hs_object_id	Long
	createdate, hs_lastmodifieddate	DateTime
product	hs_object_id	Long
	createdate, hs_lastmodifieddate	DateTime
custom_object	hs_object_id	Long
	createdate, hs_lastmodifieddate	DateTime
call	hs_object_id	Long

Entity name	Partitioning fields	Data type
	createdate, hs_lastmodifieddate	DateTime
email	hs_object_id	Long
	createdate, hs_lastmodifieddate	DateTime
meeting	hs_object_id	Long
	createdate, hs_lastmodifieddate	DateTime
note	hs_object_id	Long
	createdate, hs_lastmodifieddate	DateTime
task	hs_object_id	Long
	createdate, hs_lastmodifieddate	DateTime
postal_mail	hs_object_id	Long
	createdate, hs_lastmodifieddate	DateTime

Example:

```
hubspot_read = glueContext.create_dynamic_frame.from_options(
    connection_type="hubspot",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "company",
        "API_VERSION": "v3",
        "PARTITION_FIELD": "hs_object_id"
        "LOWER_BOUND": "50"
        "UPPER_BOUND": "16726619290"
        "NUM_PARTITIONS": "10"
    }
```

HubSpot connection options

The following are connection options for HubSpot:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your object in HubSpot.
- `API_VERSION`(String) - (Required) Used for Read. HubSpot Rest API version you want to use. For example: v1,v2,v3,v4.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.
- `TRANSFER_MODE`(String) - Used to indicate whether the query should be run on Async mode.

Limitations and notes for HubSpot connector

The following are limitations or notes for the HubSpot connector:

- The search endpoints are limited to 10,000 total results for any given query. Any partition having more than 10,000 records will result in a 400 error.
- Other important limitations for the connector are described in Limitations.
- A maximum of three filtering statements are accepted by HubSpot.
- Currently, HubSpot supports Associations between standard HubSpot objects (e.g. contact, company, deal, or ticket) and custom objects.

  - For a Free account: you can create only up to 10 association types between each object pairing (e.g. contacts and companies).
  - For a Super Admin account: you can create only up to 50 association types between each object pairing.
  - For more information, see Associations v4 and Create and use association labels.
- The 'Quote' and 'Communications' objects are not present for Associations as they are currently not supported in the connector.
- For Async, SaaS sorts the values in ascending order only.
- For the `Ticket` entity, SaaS doesn't return the `hs_object_id` field in the Async mode.

Connecting to Instagram Ads

Instagram is a popular photo-sharing app that lets you connect with brands, celebrities, thought leaders, friends, family, and more. It is a photo-sharing and social networking service. Users can take photos or short videos and share them with their followers. Instagram ads are posts for which businesses can pay to serve to Instagram users.

Topics

- AWS Glue support for Instagram Ads
- Policies containing the API operations for creating and using connections
- Configuring Instagram Ads
- Configuring Instagram Ads connections
- Reading from Instagram Ads entities
- Instagram Ads connection options
- Limitations and notes for Instagram Ads connector

AWS Glue support for Instagram Ads

AWS Glue supports Instagram Ads as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Instagram Ads.

Supported as a target?

No.

Supported Instagram Ads API versions

The following Instagram Ads API versions are supported:

- v17.0
- v18.0
- v19.0
- v20.0

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and

Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.

- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Instagram Ads

Before you can use AWS Glue to transfer data from Instagram Ads, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- Instagram Standard accounts are accessed indirectly through Facebook.
- User authentication is needed to generate the access token.
- The Instagram Ads SDK connector will be implementing the *User Access Token OAuth* flow.
- We are using OAuth2.0 to authenticate our API requests to Instagram Ads. This web-based authentication falls under the Multi-Factor Authentication (MFA) architecture, which is a superset of 2FA.
- The user needs to grant permissions to access the end points. For accessing the user's data, endpoint authorization is handled through permissions and features.

Getting OAuth 2.0 credentials

To obtain API credentials so that you can make authenticated calls to your instance, see Graph API.

Configuring Instagram Ads connections

Instagram Ads supports the AUTHORIZATION_CODE grant type for OAuth2.

- This grant type is considered three-legged OAuth as it relies on redirecting users to the third-party authorization server to authenticate the user. It is used when creating connections via the AWS Glue console.

- Users may opt to create their own connected app in Instagram Ads and provide their own client ID and client secret when creating connections through the AWS Glue console. In this scenario, they will still be redirected to Instagram Ads to login and authorize AWS Glue to access their resources.
- This grant type results in an access token. An expiring system user token is valid for 60 days from a generated or refreshed date. To create continuity, the developer should refresh the access token within 60 days. Failing to do so results in forfeiting the access token and requires the developer obtain a new one to regain API access. See Refresh Access Token.

To configure a Instagram Ads connection:

1. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Connection type**, select Instagram Ads.
   b. Select the AWS IAM role which AWS Glue can assume and has permissions for the following actions:

      JSON

      ```
      {
        "Version": "2012-10-17",
        "Statement": [
          {
            "Effect": "Allow",
            "Action": [
              "secretsmanager:DescribeSecret",
              "secretsmanager:GetSecretValue",
              "secretsmanager:PutSecretValue",
              "ec2:CreateNetworkInterface",
              "ec2:DescribeNetworkInterfaces",
              "ec2:DeleteNetworkInterface"
            ],
            "Resource": "*"
          }
        ]
      }
      ```

   c. Provide the User Managed Client Application Client ID.

   d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens. The selected secret needs to have a key `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` with the value being the Client Secret from the connected app.
   e. Select the network options if you want to use your network.
2. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Instagram Ads entities

Prerequisite

A Instagram Ads object you would like to read from. You will need the object name. The following tables shows the supported entities.

Supported entities for source:

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Campaign	Yes	Yes	No	Yes	Yes
Ad Set	Yes	Yes	No	Yes	Yes
Ads	Yes	Yes	No	Yes	Yes
Ad Creative	No	Yes	No	Yes	No
Insights - Account	No	Yes	No	Yes	No
Ad Image	Yes	Yes	No	Yes	No
Insights - Ad	Yes	Yes	No	Yes	Yes
Insights - AdSet	Yes	Yes	No	Yes	Yes
Insights - Campaign	Yes	Yes	No	Yes	Yes

Example:

```
instagramAds_read = glueContext.create_dynamic_frame.from_options(
    connection_type="instagramads",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "entityName",
        "API_VERSION": "v20.0"
    }
```

Instagram Ads entity and field details

For more information about the entities and field details see:

- Campaign
- Ad Set
- Ad
- Ad Creative
- Ad Account Insight
- Ad Image
- Ad Insights
- AdSets Insights
- Campaigns Insights

For more information, see Marketing API.

Note

Struct and List data types are converted to String data type in the response of the connectors.

Partitioning queries

You can provide the additional Spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, and `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition the query.
- `LOWER_BOUND`: an **inclusive** lower bound value of the chosen partition field.

  For the DateTime field, we accept the Spark timestamp format used in Spark SQL queries.

  Example of valid value:

  ```
  "2022-01-01T00:00:00.000Z"
  ```
- `UPPER_BOUND`: an **exclusive** upper bound value of the chosen partition field.

  Example of valid value:

  ```
  "2024-01-02T00:00:00.000Z"
  ```
- `NUM_PARTITIONS`: the number of partitions.

Example:

```
instagramADs_read = glueContext.create_dynamic_frame.from_options(
    connection_type="instagramads",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "entityName",
        "API_VERSION": "v20.0",
        "PARTITION_FIELD": "created_time"
        "LOWER_BOUND": "2022-01-01T00:00:00.000Z"
        "UPPER_BOUND": "2024-01-02T00:00:00.000Z"
        "NUM_PARTITIONS": "10"
    }
```

Instagram Ads connection options

The following are connection options for Instagram Ads:

- `ENTITY_NAME`(String) - (Required) Used for read. The name of your object in Instagram Ads.
- `API_VERSION`(String) - (Required) Used for read. Instagram Ads Graph API version you want to use. For example: v21.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for read. Columns you want to select for the object.

- `FILTER_PREDICATE`(String) - Default: empty. Used for read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for read. Number of partitions for read.

Limitations and notes for Instagram Ads connector

The following are limitations or notes for the Instagram Ads connector:

- An app's call count is the number of calls a user can make during a rolling one-hour window 200 multiplied by the number of users. For rate limit details, see Rate Limits, and Business Use Case Rate Limits.

Connecting to Intercom in AWS Glue Studio

Intercom is the Engagement OS, an open channel between your business and your customers—in product, in the moment, and on their terms—creating an ongoing dialogue that enables you to make the most of every engagement across the customer journey.

Topics

- AWS Glue support for Intercom
- Policies containing the API operations for creating and using connections
- Configuring Intercom
- Configuring Intercom connections
- Reading from Intercom entities
- Intercom connection options
- Limitations
- Creating a new Intercom account and configuring the client app

AWS Glue support for Intercom

AWS Glue supports Intercom as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Intercom.

Supported as a target?

No.

Supported Intercom API versions

v2.5. For entity support per version specific, see Reading from Intercom entities.

Policies containing the API operations for creating and using connections

The following sample policy describes the required IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

You can also use the following managed IAM policies to allow access:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, Amazon CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Intercom

Before you can use AWS Glue to transfer from Intercom, you must meet these requirements:

Minimum requirements

- You have a Intercom account. For more information, see Creating a new Intercom account and configuring the client app.
- Your Intercom account is enabled for API access.
- You should have an app under Intercom developer account that provides the client credentials that AWS Glue uses to access your data securely when it makes authenticated calls to your account. For more information, see Intercom- New Account and Client App Creation Steps.

If you meet these requirements, you’re ready to connect AWS Glue to your Intercom account.

Configuring Intercom connections

Intercom supports the `AUTHORIZATION_CODE` grant type for OAuth 2.

This grant type is considered “three-legged” OAuth as it relies on redirecting users to the third party authorization server to authenticate the user. It is used when creating connections via the AWS Glue Console. The AWS Glue Console will redirect the user to Google Ads where the user must login and allow AWS Glue the requested permissions to access their Intercom instance.

Users should provide their own client id and client secret when creating connections through the AWS Glue Console. In this scenario, they will still be redirected to Intercom to login and authorize AWS Glue to access their resources.

This grant type results in a refresh token and access token. The access token is short lived, and may be refreshed automatically without user interaction using the refresh token.

For more information on creating a connected app for Authorization Code OAuth flow, see [Ads API](#) .

To configure an Intercom connection:

1. In AWS Secrets Manager, create a secret with the following details. It is required to create a secret for each connection in AWS Glue.
    a. For customer managed connected app – Secret should contain the connected app access token, refresh token, client_id, and client_secret.
2. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
    a. When selecting a **Connection type**, select Intercom.
    b. Provide the Intercom environment.
    c. Select the IAM role which AWS Glue can assume and has permissions for following actions:

       JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

    d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.

   e. Select the network options if you want to use your network.

3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Intercom entities

Prerequisites

- An Intercom object you would like to read from. Refer to the supported entities table below to check the available entities.

Supported entities

Entity	API_Version	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Admins	v2.5	No	No	No	Yes	No
Companies	v2.5	No	Yes	No	Yes	No
Conversations	v2.5	Yes	Yes	Yes	Yes	Yes
Data Attributes	v2.5	No	No	No	Yes	No
Contacts	v2.5	Yes	Yes	Yes	Yes	Yes
Segments	v2.5	No	No	No	Yes	No
Tags	v2.5	No	No	No	Yes	No
Teams	v2.5	No	No	No	Yes	No

Example

```
Intercom_read = glueContext.create_dynamic_frame.from_options(
    connection_type="Intercom",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "company",
        "API_VERSION": "V2.5"
    }
)
```

Intercom entity and field details

Entity	Field	Data Type	Supported Operators
Admins	type	String	NA
Admins	id	String	NA
Admins	avatar	Struct	NA
Admins	name	String	NA
Admins	email	String	NA
Admins	away_mode_enabled	Boolean	NA
Admins	away_mode_reassign	Boolean	NA
Admins	has_inbox_seat	Boolean	NA
Admins	teams_ids	List	NA
Admins	job_title	String	NA
Companies	type	String	NA
Companies	id	String	NA
Companies	app_id	String	NA
Companies	created_at	DateTime	NA
Companies	remote_created_at	DateTime	NA

Entity	Field	Data Type	Supported Operators
Companies	updated_at	DateTime	NA
Companies	last_request_at	DateTime	NA
Companies	plan	Struct	NA
Companies	company_id	String	NA
Companies	name	String	NA
Companies	custom_attributes	Struct	NA
Companies	session_count	Integer	NA
Companies	monthly_spend	Integer	NA
Companies	user_count	Integer	NA
Companies	industry	String	NA
Companies	size	Integer	NA
Companies	website	String	NA
Companies	tags	Struct	NA
Companies	segments	Struct	NA
Contacts	id	String	EQUAL_TO. NOT_EQUAL_TO
Contacts	type	String	NA
Contacts	workspace_id	String	NA
Contacts	external_id	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO

Entity	Field	Data Type	Supported Operators
Contacts	role	String	EQUAL_TO. NOT_EQUAL_TO
Contacts	email	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Contacts	phone	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Contacts	name	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Contacts	avatar	String	NA
Contacts	owner_id	Integer	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN
Contacts	social_profiles	Struct	NA
Contacts	has_hard_bounced	Boolean	EQUAL_TO
Contacts	marked_email_as_spam	Boolean	EQUAL_TO
Contacts	unsubscribed_from_emails	Boolean	EQUAL_TO
Contacts	created_at	DateTime	EQUAL_TO, GREATER_THAN, LESS_THAN

Entity	Field	Data Type	Supported Operators
Contacts	updated_at	DateTime	EQUAL_TO, GREATER_THAN, LESS_THAN
Contacts	signed_up_at	DateTime	EQUAL_TO, GREATER_THAN, LESS_THAN
Contacts	last_seen_at	DateTime	EQUAL_TO, GREATER_THAN, LESS_THAN
Contacts	last_replied_at	DateTime	EQUAL_TO, GREATER_THAN, LESS_THAN
Contacts	last_contacted_at	DateTime	EQUAL_TO, GREATER_THAN, LESS_THAN
Contacts	last_email_opened_at	DateTime	EQUAL_TO, GREATER_THAN, LESS_THAN
Contacts	last_email_clicked_at	DateTime	EQUAL_TO, GREATER_THAN, LESS_THAN
Contacts	language_override	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Contacts	browser	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO

Entity	Field	Data Type	Supported Operators
Contacts	browser_version	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Contacts	browser_language	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Contacts	os	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Contacts	location	Struct	NA
Contacts	location_country	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Contacts	location_region	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Contacts	location_city	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Contacts	android_app_name	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Contacts	android_app_version	String	NA
Contacts	android_device	String	NA
Contacts	android_os_version	String	NA

Entity	Field	Data Type	Supported Operators
Contacts	android_sdk_version	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Contacts	android_last_seen_at	Date	NA
Contacts	ios_app_name	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Contacts	ios_app_version	String	NA
Contacts	ios_device	String	NA
Contacts	ios_os_version	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Contacts	ios_sdk_version	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Contacts	ios_last_seen_at	DateTime	NA
Contacts	custom_attributes	Struct	NA
Contacts	tags	Struct	NA
Contacts	notes	Struct	NA
Contacts	companies	Struct	NA
Contacts	unsubscribed_from_sms	Boolean	NA
Contacts	sms_consent	Boolean	NA

Entity	Field	Data Type	Supported Operators
Contacts	opted_out_subscription_types	Struct	NA
Contacts	referrer	String	NA
Contacts	utm_campaign	String	NA
Contacts	utm_content	String	NA
Contacts	utm_medium	String	NA
Contacts	utm_source	String	NA
Contacts	utm_term	String	NA
Conversations	type	String	NA
Conversations	id	Integer	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN
Conversations	created_at	DateTime	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN
Conversations	updated_at	DateTime	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN
Conversations	source	Struct	NA
Conversations	source_id	String	EQUAL_TO, NOT_EQUAL_TO

Entity	Field	Data Type	Supported Operators
Conversations	source_type	String	EQUAL_TO, NOT_EQUAL_TO,
Conversations	source_delivered_as	String	EQUAL_TO, NOT_EQUAL_TO,
Conversations	source_subject	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Conversations	source_body	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Conversations	source_author_id	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Conversations	source_author_type	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Conversations	source_author_name	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Conversations	source_author_email	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Conversations	source_url	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Conversations	contacts	Struct	NA

Entity	**Field**	**Data Type**	**Supported Operators**
Conversations	teammates	Struct	NA
Conversations	title	String	NA
Conversations	admin_assignee_id	Integer	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN
Conversations	team_assignee_id	Integer	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Conversations	custom_attributes	Struct	NA
Conversations	open	Boolean	EQUAL_TO
Conversations	state	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Conversations	read	Boolean	EQUAL_TO
Conversations	waiting_since	DateTime	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN
Conversations	snoozed_until	DateTime	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN
Conversations	tags	Struct	NA
Conversations	first_contact_reply	Struct	NA

Entity	Field	Data Type	Supported Operators
Conversations	priority	String	EQUAL_TO, NOT_EQUAL_TO
Conversations	topics	Struct	NA
Conversations	sla_applied	Struct	NA
Conversations	conversation_rating	Struct	NA
Conversations	conversation_rating_requested_at	DateTime	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN
Conversations	conversation_rating_replied_at	DateTime	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN
Conversations	conversation_rating_score	Integer	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN
Conversations	conversation_rating_remark	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Conversations	conversation_rating_contact_id	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Conversations	conversation_rating_admin_id	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO

Entity	Field	Data Type	Supported Operators
Conversations	statistics	Struct	NA
Conversations	statistics_time_to_assignment	Integer	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN
Conversations	statistics_time_to_admin_reply	Integer	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN
Conversations	statistics_time_to_first_close	Integer	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN
Conversations	statistics_time_to_last_close	Integer	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN
Conversations	statistics_median_time_to_reply	Integer	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN
Conversations	statistics_first_contact_reply_at	DateTime	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN

Entity	Field	Data Type	Supported Operators
Conversations	statistics_first_assignment_at	DateTime	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN
Conversations	statistics_first_admin_reply_at	DateTime	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN
Conversations	statistics_first_close_at	DateTime	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN
Conversations	statistics_last_assignment_at	DateTime	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN
Conversations	statistics_last_assignment_admin_reply_at	DateTime	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN
Conversations	statistics_last_contact_reply_at	DateTime	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN
Conversations	statistics_last_admin_reply_at	DateTime	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN

Entity	Field	Data Type	Supported Operators
Conversations	statistics_last_close_at	DateTime	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN
Conversations	statistics_last_closed_by_id	String	CONTAINS, EQUAL_TO, NOT_EQUAL_TO
Conversations	statistics_count_reopens	Integer	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN
Conversations	statistics_count_assignments	Integer	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN
Conversations	statistics_count_conversation_parts	Integer	EQUAL_TO, NOT_EQUAL_TO, GREATER_THAN, LESS_THAN
Conversations	conversation_parts	List	NA
Data Attributes	id	Integer	NA
Data Attributes	type	String	NA
Data Attributes	model	String	NA
Data Attributes	name	String	NA
Data Attributes	full_name	String	NA

Entity	Field	Data Type	Supported Operators
Data Attributes	label	String	NA
Data Attributes	description	String	NA
Data Attributes	data_type	String	NA
Data Attributes	options	List	NA
Data Attributes	api_writable	Boolean	NA
Data Attributes	ui_writable	Boolean	NA
Data Attributes	custom	Boolean	NA
Data Attributes	archived	Boolean	NA
Data Attributes	created_at	Boolean	NA
Data Attributes	updated_at	DateTime	NA
Data Attributes	admin_id	String	NA
Segments	type	String	NA
Segments	id	String	NA
Segments	name	String	NA
Segments	created_at	DateTime	NA
Segments	updated_at	DateTime	NA
Segments	person_type	String	NA
Segments	count	Integer	NA
Tags	type	String	NA
Tags	id	String	NA

Entity	Field	Data Type	Supported Operators
Tags	name	String	NA
Teams	type	String	NA
Teams	id	String	NA
Teams	name	String	NA
Teams	admin_ids	List	NA

Partitioning queries

Additional spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, `NUM_PARTITIONS` can be provided if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition query.
- `LOWER_BOUND`: an inclusive lower bound value of the chosen partition field.

  For date, we accept the Spark date format used in Spark SQL queries. Example of valid values: `"2024-02-06"`.
- `UPPER_BOUND`: an exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: number of partitions.

Entity-wise partitioning field support details are captured in the following table.

Entity Name	Partitioning Field	Data Type
Contacts	created_at, updated_at,last_seen_at	DateTime
Conversations	id	Integer
Conversations	created_at, updated_at	DateTime

Example

```
Intercom_read = glueContext.create_dynamic_frame.from_options(
    connection_type="Intercom",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "conversation",
        "API_VERSION": "V2.5",
        "PARTITION_FIELD": "created_at"
        "LOWER_BOUND": "2022-07-13T07:55:27.065Z"
        "UPPER_BOUND": "2022-08-12T07:55:27.065Z"
        "NUM_PARTITIONS": "2"
    }
)
```

Intercom connection options

The following are connection options for Intercom:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your Object in Intercom.
- `API_VERSION`(String) - (Required) Used for Read. Intercom Rest API version you want to use. Example: v2.5.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.
- `INSTANCE_URL`(String) - URL of the instance where the user wants to run the operations. For example: https://api.intercom.io.

Limitations

The following are limitations for the Intercom connector:

- When using the Company entity, there is a limit of 10,000 Companies that can be returned. For more information, see List all companies API.
- While applying order by, filter is mandatory for both **Contact** and **Conversation** entities.
- MCA is supported by the SaaS provider. However, based on the API rate limits mentioned in the documentation, we will not host MCA on AWS Glue as it may impact other workloads and potentially cause performance issues due to resource contention.

Creating a new Intercom account and configuring the client app

Creating a Intercom account

1. Choose on the Intercom URL and choose **Start my free trial** on right upper corner of the page.
2. Choose **Try for free button** on right upper corner of the page.
3. Choose the business type you require.
4. Enter all the information required on the page.
5. After entering all the information, choose **Register**.

Creating an Intercom developer app

To get the **Client Id** and **Client Secret**, you create a developer account.

1. Navigate to https://app.intercom.com/.
2. Enter the Email ID and Password/ Sign In Using Google and log in.
3. Choose **user profile** on the left bottom corner and choose settings.
4. Choose **Apps & Integration**.
5. Choose the **Developer Hub** tab under **Apps & Integration**.
6. Choose **New app** and create the app here.
7. Provide the app name and choose **Create** app.
8. Inside the app, navigate to the **Authentication** section.

9. Choose the **edit** and add redirect URIs. Add the your region-specific Redirect URL as `https://<aws-region>.console.aws.amazon.com/gluestudio/oauth`. For example, add `https://us-east-1.console.aws.amazon.com/gluestudio/oauth for the us-east-1 region`.
10. Get the generated **Client Id** and **Client Secret** in the Basic Information Section.

Connecting to Jira Cloud

Jira Cloud is a platform developed by Atlassian. The platform includes issue tracking products that help teams plan and track their Agile projects. As a Jira Cloud user, your account contains data about your projects, such as issues, workflows, and events. You can use AWS Glue to transfer your Jira Cloud data to certain AWS services or other supported applications.

Topics

- AWS Glue support for Jira Cloud
- Policies containing the API operations for creating and using connections
- Configuring Jira Cloud
- Configuring Jira Cloud connections
- Reading from Jira Cloud entities
- Jira Cloud connection options
- Limitations and notes for Jira Cloud connector

AWS Glue support for Jira Cloud

AWS Glue supports Jira Cloud as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Jira Cloud.

Supported as a target?

No.

Supported Jira Cloud API versions

The following Jira Cloud API versions are supported:

- v2

For entity support per version specific, see Supported entities for Source.

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention

for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Jira Cloud

Before you can use AWS Glue to transfer data from Jira Cloud to supported destinations, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- You have an Atlassian account where you use the Jira software product in Jira Cloud. For more information, see Creating a Jira Cloud account.
- You must have an AWS account created with the service access to AWS Glue.
- This app provides the client credentials that AWS Glue uses to access your data securely when it makes authenticated calls to your account. For more information, see Enabling OAuth 2.0 (3LO) in the Atlassian Developer documentation.

If you meet these requirements, you're ready to connect AWS Glue to your Jira Cloud account.

Creating a Jira Cloud account

To create a Jira Cloud account:

1. Navigate to the Atlassian sign up URL.
2. Enter your work email and name and choose **Agree**. You receive a verification email.
3. After verifying your email, you can create a password and choose **Sign up**.
4. Enter name and password, and choose **Sign up**.
5. You are redirected to a page where you need to enter your site. Enter a site name and choose **Agree**.

Once your Atlassian Cloud site starts up, you can set up your Jira by answering a few questions as per your project type preferences.

To log in to an existing account:

1. Navigate to the Atlassian login URL and enter credentials.

2. Enter email and password and click **Log in**. You are redirected to the Jira dashboard.

Creating an app in your Jira Cloud

To create an app in Jira Cloud and obtain the Client ID and Client Secret from the managed client app:

1. Navigate to the Jira Cloud URL and enter credentials.
2. Choose **Create** and select the **OAuth 2.0 integration** option.
3. Enter the app name, check **T&C** and choose **Create**.
4. Navigate to the **Distribution** section in the left menu and choose **Edit**.
5. In the **Edit distribution controls** section:
    a. Select **DISTRIBUTION STATUS** as **Sharing**.
    b. Enter the vendor name.
    c. Enter the URL for your **Privacy policy**. For example, https://docs.aws.amazon.com/glue/latest/dg/security-iam-awsmanpol.html
    d. Enter the URL for your **Terms of service** (optional).
    e. Enter the URL for your **Customer support contact** (optional).
    f. Select Yes/No from the **PERSONAL DATA DECLARATION** and choose **Save changes**.
6. Navigate to **Permissions** in the left menu for the respective app.
7. For **Jira API**, choose **Add**. Once added, choose the **Configuration** option.
8. Under the **Classic scopes** > **Jira platform REST API** section choose **Edit Scopes**. and check all scopes. Click **Save**.
9. Under **Granular Scopes** choose **Edit Scopes** and select the following scopes:
10. Scroll down and find scopes. There are two kinds of scopes you must select under headings "CRM" and "Standard".
11. Add the following scopes:

```
read:application-role:jira
read:audit-log:jira
read:avatar:jira
read:field:jira
read:group:jira
read:instance-configuration:jira
```

```
read:issue-details:jira
read:issue-event:jira
read:issue-link-type:jira
read:issue-meta:jira
read:issue-security-level:jira
read:issue-security-scheme:jira
read:issue-type-scheme:jira
read:issue-type-screen-scheme:jira
read:issue-type:jira
read:issue.time-tracking:jira
read:label:jira
read:notification-scheme:jira
read:permission:jira
read:priority:jira
read:project:jira
read:project-category:jira
read:project-role:jira
read:project-type:jira
read:project-version:jira
read:project.component:jir
read:project.property:jira
read:resolution:jira
read:screen:jira
read:status:jira
read:user:jira
read:workflow-scheme:jira
read:workflow:jira
read:field-configuration:jira
read:issue-type-hierarchy:jira
read:webhook:jira
```

12. Navigate to **Authentication** in the left menu and choose **Add**.
13. Enter a **Callback URL** such as https://us-east-1.console.aws.amazon.com/gluestudio/oauth
14. Navigate to **Settings** in the left menu and scroll down for **Authentication** details. Note the Client ID and Secret.

Configuring Jira Cloud connections

Jira Cloud supports the AUTHORIZATION_CODE grant type for OAuth2.

- This grant type is considered "three-legged" OAuth as it relies on redirecting users to a third-party authorization server to authenticate the user. It is used when creating connections via the

AWS Glue console. The AWS Glue console will redirect the user to Jira Cloud where the user must login and allow AWS Glue the requested permissions to access their Jira Cloud instance.

- Users may still opt to create their own connected app in Jira Cloud and provide their own client ID and client secret when creating connections through the AWS Glue console. In this scenario, they will still be redirected to Jira Cloud to login and authorize AWS Glue to access their resources.
- This grant type results in a refresh token and access token. The access token is short lived, and may be refreshed automatically without user interaction using the refresh token.
- For public Jira Cloud documentation on creating a connected app for Authorization Code OAuth flow, see Enabling OAuth 2.0 (3LO).

To configure a Jira Cloud connection:

1. In AWS Secrets Manager, create a secret with the following details:
   a. For the customer managed connected app, the Secret should contain the connected app Consumer Secret with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as key.
   b. Note: You must create a secret for the connection in AWS Glue.
2. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Connection type**, select Jira Cloud.
   b. Provide the Jira Cloud environment.
   c. Select the AWS IAM role which AWS Glue can assume and has permissions for following actions:

      JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

   d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.

   e. Select the network options if you want to use your network.

3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Jira Cloud entities

Prerequisite

A Jira Cloud object you would like to read from. You will need the object name such as Audit Record or Issue. The following table shows the supported entities.

Supported entities for source:

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Audit Record	Yes	Yes	No	Yes	Yes
Issue	Yes	Yes	No	Yes	Yes
Issue Field	No	No	No	Yes	No
Issue Field Configuration	Yes	Yes	No	Yes	Yes
Issue Link Type	No	No	No	Yes	No
Issue Notification Scheme	Yes	Yes	No	Yes	Yes

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Issue Security Scheme	No	No	No	Yes	No
Issue Type Scheme	Yes	Yes	Yes	Yes	Yes
Issue Type Screen Scheme	Yes	Yes	Yes	Yes	Yes
Issue Type	No	No	No	Yes	No
Jira Setting	Yes	No	No	Yes	No
Jira Setting Advanced	No	No	No	Yes	No
Jira Setting Global	No	No	No	Yes	No
Label	No	No	No	Yes	Yes
Myself	Yes	No	No	Yes	No
Permission	No	No	No	Yes	No.
Project	Yes	Yes	Yes	Yes	Yes
Project Category	No	No	No	Yes	No
Project Type	No	No	No	Yes	No
Server Info	No	No	No	Yes	No
Users	No	No	No.	Yes	No
Workflow	Yes	Yes	Yes	Yes	Yes

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Workflow Scheme	No	Yes	No	Yes	Yes
Workflow Scheme Project Association	Yes	No	No	Yes	No
Workflow Status	No	No	No	Yes	No
Workflow Status Category	No	No	No	Yes	No

Example:

```
jiracloud_read = glueContext.create_dynamic_frame.from_options(
    connection_type="JiraCloud",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "audit-record",
        "API_VERSION": "v2"
    }
```

Jira Cloud entity and field details:

Object	Field	Data type	Filter operators supported
Audit Record	filter	String	"="
	from	DateTime	"="
	to	DateTime	"="

Object	Field	Data type	Filter operators supported
	id	Integer	N/A
	summary	String	N/A
	remoteAddress	String	N/A
	authorAccountId	String	N/A
	created	String	N/A
	category	String	N/A
	eventSource	String	N/A
	description	String	N/A
	objectItem	Struct	N/A
	changedValues	List	N/A
	associatedItems	List	N/A
Groups	groupName	List	"="
	name	String	N/A
	groupId	String	"="
Issue	affectedVersion	String	"=, !="
	assignee	String	"=, !="
	category	String	"=, !="
	component	String	"=, !="
	creator	String	"=, !="
	due	DateTime	N/A

Object	Field	Data type	Filter operators supported
	epic_link	String	"=, !="
	filter	String	"=, !="
	fixVersion	String	"=, !="
	hierarchyLevel	Integer	"=, !="
	issueKey	String	"=, !=, >, <, >=, <="
	issueLink	String	"=, !="
	issueLinkType	String	"=, !="
	labels	String	"=, !="
	lastViewed	DateTime	"=, >, <, >=, <=, between"
	level	String	"=, !="
	parent	String	"=, !="
	priority	String	"=, !="
	project	String	"=, !="
	projectType	String	"=, !="
	reporter	String	"=, !="
	resolution	String	"=, !="
	resolved	DateTime	"=, >, <, >=, <=, between"
	sprint	String	"=, !="
	status	String	"=, !="

Object	Field	Data type	Filter operators supported
	type	String	"=, !="
	updated	DateTime	"=, >, <, >=, <=, between"
	voter	String	"=, !="
	votes	Integer	"=, !=, <, >, <=, >=, between"
	watcher	String	"=, !="
	watchers	Integer	"=, !=, <, >, <=, >=, between"
	workRatio	Integer	"=, !=, <, >, <=, >=, between"
	validateQuery	String	"="
	expand	String	"="
	fieldByKeys	Boolean	"="
	id	String	N/A
	self	String	N/A
	key	String	N/A
	renderedFields	Struct	N/A
	properties	List	"="
	names	Struct	N/A
	schema	Struct	N/A

Object	Field	Data type	Filter operators supported
	transitions	List	N/A
	operations	Struct	N/A
	editmeta	Struct	N/A
	changelog	Struct	N/A
	versionedRepresentations	Struct	N/A
	fields	List	"="
	fieldsToInclude	Struct	N/A
	warningMessages	List	N/A
	created	DateTime	N/A
	worklogDate	DateTime	N/A
IssueEvents	id	Integer	N/A
	name	String	N/A
Issue Fields	id	String	N/A
	key	String	N/A
	name	String	N/A
	custom	Boolean	N/A
	orderable	Boolean	N/A
	navigable	Boolean	N/A
	searchable	Boolean	N/A

Object	Field	Data type	Filter operators supported
	clauseNames	List	N/A
	scope	Struct	N/A
	schema	Struct	N/A
Issue Field Configurations	isDefault	Boolean	"="
	query	String	"="
	id	Integer	"="
	name	String	N/A
	description	String	N/A
Issue Link Type	id	String	N/A
	name	String	N/A
	inward	String	N/A
	outward	String	N/A
	self	String	N/A
Issue Notification Schemes	expand	String	"="
	self	String	N/A
	id	Integer	N/A
	name	String	N/A
	description	String	N/A
	notificationSchemeEvents	List	N/A

Object	Field	Data type	Filter operators supported
	scope	Struct	N/A
Issue Priority	self	String	N/A
	statusColor	String	N/A
	description	String	N/A
	iconUrl	String	N/A
	name	String	N/A
	id	String	N/A
	isDefault	Boolean	N/A
Issue Resolutions	self	String	N/A
	id	String	N/A
	description	String	N/A
	name	String	N/A
Issue Security Scheme	self	String	N/A
	id	Integer	N/A
	name	String	N/A
	description	String	N/A
	defaultSecurityLevelId	Integer	N/A
	levels	List	N/A
Issue Type	self	String	N/A

Object	Field	Data type	Filter operators supported
	id	String	N/A
	description	String	N/A
	iconUrl	String	N/A
	name	String	N/A
	subtask	Boolean	N/A
	avatarId	Integer	N/A
	entityId	String	N/A
	hierarchyLevel	Integer	N/A
	scope	Struct	N/A
Issue Type Scheme	orderBy	String	"="
	expand	String	"="
	queryString	String	"="
	id	String	N/A
	name	String	N/A
	description	String	N/A
	defaultIssueTypeId	String	N/A
	isDefault	Boolean	N/A
Issue Type Screen Scheme	queryString	String	"="
	orderBy	String	"="
	expand	String	"="

Object	Field	Data type	Filter operators supported
	id	String	"="
	name	String	N/A
	description	String	N/A
Jira Settings	key	String	N/A
	keyFilter	String	"="
	id	String	N/A
	value	String	N/A
	name	String	N/A
	desc	String	N/A
	type	String	N/A
	defaultValue	String	N/A
	example	String	N/A
	allowedValues	List	N/A
Jira Settings Advanced	id	String	N/A
	key	String	N/A
	value	String	N/A
	name	String	N/A
	desc	String	N/A
	type	String	N/A
	defaultValue	String	N/A

Object	Field	Data type	Filter operators supported
	example	String	N/A
	allowedValues	List	N/A
Jira Settings Global	votingEnabled	Boolean	N/A
	watchingEnabled	Boolean	N/A
	unassignedIssuesAllowed	Boolean	N/A
	subTasksEnabled	Boolean	N/A
	issueLinkingEnabled	Boolean	N/A
	timeTrackingEnabled	Boolean	N/A
	attachmentsEnabled	Boolean	N/A
	timeTrackingConfiguration	Struct	N/A
Label	values	List	N/A
Myself	expand	String	"="
	self	String	N/A
	accountId	String	N/A
	accountType	String	N/A
	emailAddress	String	N/A
	avatarUrls	String	N/A
	displayName	String	N/A
	active	Boolean	N/A

Object	Field	Data type	Filter operators supported
	timeZone	String	N/A
	locale	String	N/A
	groups	Struct	N/A
	applicationRoles	Struct	N/A
Permission	id	String	N/A
	key	String	N/A
	name	String	N/A
	type	String	N/A
	description	String	N/A
	havePermission	Boolean	N/A
	deprecatedKey	Boolean	N/A
Project	orderBy	String	"="
	keys	List	"="
	query	String	"="
	typeKey	String	"="
	categoryId	Integer	"="
	action	String	"="
	expand	String	"="
	status	List	"="
	self	String	N/A

Object	Field	Data type	Filter operators supported
	id	Integer	"="
	key	String	N/A
	description	String	N/A
	lead	Struct	N/A
	components	List	N/A
	issueTypes	List	N/A
	url	String	N/A
	email	String	N/A
	assigneeType	String	N/A
	versions	List	N/A
	name	String	N/A
	roles	Struct	N/A
	avatarUrls	Struct	N/A
	projectCategory	Struct	N/A
	projectTypeKey	String	N/A
	simplified	Boolean	N/A
	style	String	N/A
	favourite	Boolean	N/A
	isPrivate	Boolean	N/A
	issueTypeHierarchy	Struct	N/A

Object	Field	Data type	Filter operators supported
	permissions	Struct	N/A
	properties	List	"="
	uuid	String	N/A
	insight	Struct	N/A
	deleted	Boolean	N/A
	retentionTillDate	String	N/A
	deletedDate	String	N/A
	deletedBy	Struct	N/A
	archived	Boolean	N/A
	archivedDate	String	N/A
	archivedBy	Struct	N/A
	landedPageInfo	Struct	N/A
Project Category	self	String	N/A
	id	String	N/A
	name	String	N/A
	description	String	N/A
Project Type	key	String	N/A
	formattedKey	String	N/A
	description	String	N/A
	descriptionI18nKey	String	N/A

Object	Field	Data type	Filter operators supported
	icon	String	N/A
	color	String	N/A
Server Info	baseUrl	String	N/A
	version	String	N/A
	versionNumbers	List	N/A
	deploymentType	String	N/A
	buildNumber	Integer	N/A
	buildDate	DateTime	N/A
	serverTime	DateTime	N/A
	scmInfo	String	N/A
	serverTitle	String	N/A
	healthChecks	List	N/A
Users	self	String	N/A
	accountId	String	N/A
	accountType	String	N/A
	emailAddress	String	N/A
	avatarUrls	Struct	N/A
	displayName	String	N/A
	active	Boolean	N/A
	timeZone	String	N/A

Object	Field	Data type	Filter operators supported
	locale	String	N/A
	groups	Struct	N/A
	applicationRoles	Struct	N/A
	expand	String	N/A
Workflow	workflowName	String	"="
	expand	String	"="
	queryString	String	"="
	orderBy	String	"="
	isActive	Boolean	"="
	id	Struct	N/A
	description	String	N/A
	transitions	List	N/A
	statuses	List	N/A
	isDefault	Boolean	N/A
	schemes	List	N/A
	projects	List	N/A
	hasDraftWorkflow	Boolean	N/A
	operations	Struct	N/A
	created	String	N/A
	updated	String	N/A

Object	Field	Data type	Filter operators supported
Workflow Scheme	self	String	N/A
	id	Integer	N/A
	name	String	N/A
	description	String	N/A
	defaultWorkflow	String	N/A
	issueTypeMappings	Struct	N/A
	originalDefaultWorkflow	String	N/A
	originalIssueTypeMappings	Struct	N/A
	draft	Boolean	N/A
	lastModifiedUser	Struct	N/A
	lastModified	String	N/A
	updateDraftIfNeeded	Boolean	N/A
	issueTypes	Struct	N/A
Workflow Scheme Project Association	projectId	Integer	"="
	projectIds	List	N/A
	workflowScheme	Struct	N/A
Workflow Status	self	String	N/A
	description	String	N/A
	iconUrl	String	N/A

Object	Field	Data type	Filter operators supported
	name	String	N/A
	id	String	N/A
	StatusCategory	Struct	N/A
Workflow Status Category	self	String	N/A
	id	String	N/A
	key	String	N/A
	colorName	String	N/A
	name	String	N/A

Partitioning queries

You can provide the additional Spark option `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With this parameter, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

- `NUM_PARTITIONS`: the number of partitions.

Example:

```
jiraCloud_read = glueContext.create_dynamic_frame.from_options(
    connection_type="JiraCloud",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "issue",
        "API_VERSION": "v2",
        "NUM_PARTITIONS": "10"
    }
```

Jira Cloud connection options

The following are connection options for Jira Cloud:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your object in Jira Cloud.
- `API_VERSION`(String) - (Required) Used for Read. Jira Cloud Rest API version you want to use. For example: v2.
- `DOMAIN_URL`(String) - (Required) The Jira Cloud ID you want to use.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.

Limitations and notes for Jira Cloud connector

The following are limitations or notes for the Jira Cloud connector:

- The `Contains` operator does not work with the `resourceName` field, which is of String data type.

Connecting to Kustomer

Kustomer is a powerful customer experience platform that brings together everything you need to serve your customers better in one easy-to-use tool.

Topics

- AWS Glue support for Kustomer
- Policies containing the API operations for creating and using connections
- Configuring Kustomer
- Configuring Kustomer connections
- Reading from Kustomer entities
- Kustomer connection options
- Kustomer limitations

AWS Glue support for Kustomer

AWS Glue supports Kustomer as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Kustomer.

Supported as a target?

No.

Supported Kustomer API versions

The following Kustomer API versions are supported:

- v1

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Kustomer

Before you can use AWS Glue to transfer data from Kustomer to supported destinations, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- You have an account with Kustomer that contains the data that you want to transfer.
- In the settings for your account, you've created a API key. For more information, see Creating an API key.
- You provide the API key to AWS Glue while creating the connection.

If you meet these requirements, you’re ready to connect AWS Glue to your Kustomer account.

Creating an API key

To create an API key that you will use to create a connection for the Kustomer connector in AWS Glue Studio:

1. Log in to the Kustomer dashboard using your credentials.
2. Choose the **Settings** icon from the left menu.
3. Expand the **Security** drop down and select **API Keys**.
4. In the API Key creation page select **Add an API Key** from the top right corner.
5. Fill the mandatory inputs for the API key being created.

  - Name: any name for your API Key.
  - Roles: 'org' must be selected for the Kustomer APIs to function.
  - Expires (in days): the number of days you want the API key to be valid. You can keep it as **Never expires**, if it suits your use case.

6. Choose **Create**.
7. Store the API key (token) value for further usage to create a connection for the Kustomer connector in AWS Glue Studio.

Configuring Kustomer connections

To configure a Kustomer connection:

1. In AWS Secrets Manager, create a secret with the following details:
   a. For the customer managed connected app, the Secret should contain the connected app Consumer Secret with `apiKey` as the key.
   b. Note: you must create a secret for your connections in AWS Glue.

1. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. Under **Connections**, choose **Create connection**.
   b. When selecting a **Data Source**, select Kustomer.
   c. Select the AWS IAM role which AWS Glue can assume and has permissions for following actions:

      JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
            "ec2:DeleteNetworkInterface"
          ],
          "Resource": "*"
        }
      ]
    }
```

   d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.
   e. Select the network options if you want to use your network.

2. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Kustomer entities

Prerequisite

A Kustomer object you would like to read from. You will need the object name such as Brands or Cards. The following table shows the supported entities.

Supported entities for source:

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Brands	No	Yes	No	Yes	No
Cards	No	Yes	No	Yes	No
Chat Settings	No	No	No	Yes	No
Companies	Yes	Yes	Yes	Yes	Yes
Conversations	Yes	Yes	Yes	Yes	Yes
Customers	Yes	Yes	Yes	Yes	Yes
Customer Searches Pinned	No	Yes	No	Yes	No

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Customer Searches Position	No	No	No	Yes	No
Email Hooks	No	Yes	No	Yes	No
Web Hooks	No	Yes	No	Yes	No
KB Articles	No	Yes	No	Yes	No
KB Categories	No	Yes	No	Yes	No
KB Forms	No	Yes	No	Yes	No
KB Routes	No	Yes	No	Yes	No
KB Tags	No	Yes	No	Yes	No
KB Templates	No	Yes	No	Yes	No
KB Themes	No	Yes	No	Yes	No
Klasses	No	Yes	No	Yes	No
KViews	No	Yes	No	Yes	No
Messages	Yes	Yes	Yes	Yes	Yes
Notes	Yes	Yes	Yes	Yes	Yes
Notifications	No	Yes	No	Yes	No

Example:

```
Kustomer_read = glueContext.create_dynamic_frame.from_options(
    connection_type="kustomer",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "brands",
        "API_VERSION": "v1"
    }
```

Kustomer entity and field details

For more information about the entities and field details see:

- Brands
- Cards
- Chat Settings
- Companies
- Conversations
- Customers
- Customers Searches Pinned
- Customer Searches Positions
- Hooks Email
- Hooks Web
- KB Articles
- KB Categories
- KB Forms
- KB Routes
- KB Tags
- KB Templates
- KB Themes
- Klasses
- Kviews
- Messages
- Notes
- Notifications

Kustomer API v1

Entity	Field	Data type	Supported operators
Brands	id	String	N/A
	name	String	N/A
	iconUrl	String	N/A
	createdAt	DateTime	N/A
	updatedAt	DateTime	N/A
	modifiedAt	DateTime	N/A
	default	Boolean	N/A
Cards	id	String	N/A
	name	String	N/A
	createdAt	DateTime	N/A
	updatedAt	DateTime	N/A
	modifiedAt	DateTime	N/A
	description	String	N/A
	url	String	N/A
	contexts	List	N/A
Chat Settings	id	String	N/A
	settingsVersion	Integer	N/A
	widgetType	String	N/A
	version	Integer	N/A
	teamName	String	N/A

Entity	Field	Data type	Supported operators
	greeting	String	N/A
	autoreply	String	N/A
	embedIconUrl	String	N/A
	embedIconColor	String	N/A
	fallbackEmailSubject	String	N/A
	fallbackEmailIntroduction	String	N/A
	enabled	Boolean	N/A
	outboundChatEnabled	Boolean	N/A
	updatedAt	DateTime	N/A
	modifiedAt	DateTime	N/A
	offhoursMessage	String	N/A
	offhoursImageUrl	String	N/A
	closableChat	Boolean	N/A
	noHistory	Boolean	N/A
	disableAttachments	Boolean	N/A
	volumeControl	Struct	N/A
	singleSessonChat	Boolean	N/A
	showTypingIndicatorWeb	Boolean	N/A
Companies	id	String	N/A

Entity	Field	Data type	Supported operators
	name	String	=, !=, CONTAINS
	createdAt	DateTime	N/A
	updatedAt	DateTime	N/A
	modifiedAt	DateTime	=, !=, >, >=, <, <=, BETWEEN
	tags	List	N/A
	domains	List	N/A
	emails	List	N/A
	phones	List	N/A
	whatsapps	List	N/A
	socials	List	N/A
	urls	List	N/A
	locations	List	N/A
	roleGroupVersions	List	N/A
	rev	Integer	N/A
Conversations	id	String	N/A
	name	String	=, !=, CONTAINS
	preview	String	N/A
	channels	List	N/A
	status	String	=, !=, CONTAINS
	messageCount	Integer	=, !=, >, >=, <, <=

Entity	Field	Data type	Supported operators
	noteCount	Integer	=, !=, >, >=, <, <=
	satisfaction	Integer	=, !=, >, >=, <, <=
	satisfactionLevel	Struct	N/A
	createdAt	DateTime	=, !=, >, >=, <, <=, BETWEEN
	updatedAt	DateTime	=, !=, >, >=, <, <=, BETWEEN
	modifiedAt	DateTime	=, !=, >, >=, <, <=, BETWEEN
	lastActivityAt	DateTime	N/A
	spam	Boolean	N/A
	ended	Boolean	=, !=
	endedAt	DateTime	=, !=, >, >=, <, <=, BETWEEN
	endedReason	String	CONTAINS
	endedByType	String	N/A
	importedAt	String	N/A
	tags	List	N/A
	suggestedTags	List	N/A
	sentiment	String	N/A
	predictions	List	N/A
	suggestedShortcuts	List	N/A

Entity	Field	Data type	Supported operators
	firstMessageIn	Struct	N/A
	firstMessageOut	Struct	N/A
	lastMessageIn	Struct	N/A
	lastMessageOut	Struct	N/A
	lastMessageAt	DateTime	=, !=, >, >=, <, <=, BETWEEN
	lastMessageUnrespondedTo	Struct	N/A
	lastMessageUnrespondedToSinceLastDone	Struct	N/A
	assignedUsers	List	N/A
	assignedTeams	List	N/A
	firstResponse	Struct	N/A
	firstResponseSinceLastDone	Struct	N/A
	lastResponse	Struct	N/A
	firstDone	Struct	N/A
	lastDone	Struct	N/A
	direction	String	=, !=, CONTAINS
	lastMessageDirection	String	N/A
	outboundMessageCount	Integer	N/A

Entity	Field	Data type	Supported operators
	inboundMessageCount	Integer	N/A
	rev	Integer	N/A
	priority	Integer	=, !=, >, >=, <, <=
	roleGroupVersions	List	N/A
	accessOverride	List	N/A
	assistant	Struct	N/A
	phase	String	N/A
	Skills	List	N/A
	matchedTimeBasedRules	List	N/A
Customers	id	String	N/A
	name	String	=, !=, CONTAINS
	displayName	String	N/A
	displayColor	String	N/A
	displayIcon	String	N/A
	externalId	String	=, !=, CONTAINS
	externalIds	List	N/A
	sharedExternalIds	List	N/A
	emails	List	N/A
	sharedEmails	List	N/A

Entity	Field	Data type	Supported operators
	phones	List	N/A
	sharedPhones	List	N/A
	whatsapps	List	N/A
	facebookIds	List	N/A
	instagramIds	List	N/A
	socials	List	N/A
	sharedSocials	List	N/A
	urls	List	N/A
	locations	List	N/A
	activeUsers	List	N/A
	watchers	List	N/A
	recentLocation	Struct	N/A
	locale	String	=, !=, CONTAINS
	timeZone	String	N/A
	gender	String	=, !=, CONTAINS
	createdAt	DateTime	=, !=, >, >=, <, <=, BETWEEN
	updatedAt	DateTime	=, !=, >, >=, <, <=, BETWEEN
	modifiedAt	DateTime	=, !=, >, >=, <, <=, BETWEEN
	lastActivityAt	DateTime	N/A

Entity	Field	Data type	Supported operators
	deleted	Boolean	N/A
	lastConversation	Struct	N/A
	conversationCounts	Struct	N/A
	preview	Struct	N/A
	tags	List	N/A
	progressiveStatus	String	=, !=, CONTAINS
	verified	Boolean	N/A
	rev	Integer	N/A
	recentItems	List	N/A
	defaultLang	String	=, !=, CONTAINS
	satisfactionLevel	Struct	N/A
	roleGroupVersions	List	N/A
	accessOverride	List	N/A
	companyName	String	N/A
	firstName	String	N/A
	lastName	String	N/A
Customer Searches Pinned	id	String	N/A
	search	String	N/A
	createdAt	DateTime	N/A
Customer Searches Positions	id	String	N/A

Entity	Field	Data type	Supported operators
	positions	List	N/A
	children	List	N/A
	createdAt	DateTime	N/A
	updatedAt	DateTime	N/A
	modifiedAt	DateTime	N/A
	rev	Integer	N/A
Email Hooks	id	String	N/A
	description	String	N/A
	debug	Boolean	N/A
	email	String	N/A
	eventName	String	N/A
	title	String	N/A
	hash	String	N/A
	key	String	N/A
	createdAt	DateTime	N/A
	modifiedAt	DateTime	N/A
	updatedAt	DateTime	N/A
Web Hooks	id	String	N/A
	description	String	N/A
	eventName	String	N/A

Entity	Field	Data type	Supported operators
	hash	String	N/A
	url	String	N/A
	createdAt	DateTime	N/A
	modifiedAt	DateTime	N/A
	updatedAt	DateTime	N/A
	title	String	N/A
	version	Integer	N/A
	debug	Boolean	N/A
KB Articles	id	String	N/A
	hash	String	N/A
	title	String	N/A
	source	String	N/A
	status	String	N/A
	scope	String	N/A
	createdAt	DateTime	N/A
	updatedAt	DateTime	N/A
	deleted	Boolean	N/A
	deletedAt	DateTime	N/A
	modifiedAt	DateTime	N/A
	publishedAt	DateTime	N/A

Entity	Field	Data type	Supported operators
	tags	List	N/A
	categories	List	N/A
	knowledgeBases	List	N/A
	metaTitle	String	N/A
	metaDescription	String	N/A
	metaKeywords	List	N/A
	langVersions	Struct	N/A
	latestLangs	Struct	N/A
KB Categories	id	String	N/A
	hash	String	N/A
	createdAt	DateTime	N/A
	modifiedAt	DateTime	N/A
	updatedAt	DateTime	N/A
	published	Boolean	N/A
	positions	List	N/A
	categoryPositions	List	N/A
	root	Boolean	N/A
	langs	Struct	N/A
KB Forms	id	String	N/A
	name	String	N/A

Entity	Field	Data type	Supported operators
	slug	String	N/A
	hash	String	N/A
	body	String	N/A
	layout	List	N/A
	layoutV2	List	N/A
	componentsV2	Struct	N/A
	conditions	Struct	N/A
	advanced	Boolean	N/A
	createdAt	DateTime	N/A
	updatedAt	DateTime	N/A
	publishedAt	DateTime	N/A
	modifiedAt	String	N/A
	published	Boolean	N/A
	snippets	List	N/A
	recaptcha	Boolean	N/A
	klass	String	N/A
	channel	String	N/A
	deflection	Boolean	N/A
	formHookEnabled	Boolean	N/A
	replyFrom	String	N/A

Entity	Field	Data type	Supported operators
	wcag	Boolean	N/A
KB Routes	id	String	N/A
	url	String	N/A
	routableType	String	N/A
	routableId	String	N/A
	createdAt	DateTime	N/A
	updatedAt	DateTime	N/A
	modifiedAt	DateTime	N/A
KB Tags	id	String	N/A
	name	String	N/A
	createdAt	DateTime	N/A
	updatedAt	DateTime	N/A
	modifiedAt	DateTime	N/A
KB Templates	id	String	N/A
	title	String	N/A
	description	String	N/A
	beta	Boolean	N/A
	manifest	Struct	N/A
	jsxSnippets	List	N/A
	images	List	N/A

Entity	Field	Data type	Supported operators
	version	String	N/A
	createdAt	DateTime	N/A
	updatedAt	DateTime	N/A
KB Themes	id	String	N/A
	name	String	N/A
	active	Boolean	N/A
	default	Boolean	N/A
	lastfileUpdatedAt	DateTime	N/A
	custom	Boolean	N/A
	status	String	N/A
	templateVersionId	String	N/A
	templateTitle	String	N/A
	templateVersion	String	N/A
	manifest	Struct	N/A
	configSnippets	List	N/A
	jsxSnippets	List	N/A
	createdAt	DateTime	N/A
	updatedAt	DateTime	N/A
	modifiedAt	DateTime	N/A
	rev	Integer	N/A

Entity	Field	Data type	Supported operators
Klasses	id	String	N/A
	name	String	N/A
	icon	String	N/A
	color	String	N/A
	appDisabled	Boolean	N/A
	status	String	N/A
	updatedAt	DateTime	N/A
	createdAt	DateTime	N/A
	s3DataUrl	String	N/A
KViews	id	String	N/A
	resource	String	N/A
	template	String	N/A
	context	String	N/A
	meta	Struct	N/A
	appDisabled	Boolean	N/A
	enabled	Boolean	N/A
	advanced	Boolean	N/A
	layout	List	N/A
	components	Struct	N/A
	conditions	Struct	N/A

Entity	Field	Data type	Supported operators
	rev	Integer	N/A
	createdAt	DateTime	N/A
	modifiedAt	DateTime	N/A
	updatedAt	DateTime	N/A
Notifications	id	String	N/A
	name	String	N/A
	status	String	N/A
	event	Struct	N/A
	createdAt	DateTime	N/A
	updatedAt	DateTime	N/A
Messages	id	String	N/A
	externalId	String	N/A
	channel	String	=, !=, CONTAINS
	app	String	N/A
	size	Integer	=, !=, >, >=, <, <=
	direction	String	=, !=, CONTAINS
	preview	String	N/A
	subject	String	N/A
	meta	Struct	N/A
	status	String	=, !=, CONTAINS

Entity	Field	Data type	Supported operators
	directionType	String	=, !=, CONTAINS
	assignedTeams	List	N/A
	assignedUsers	List	N/A
	errorAt	DateTime	=, !=, >, >=, <, <=, BETWEEN
	auto	Boolean	=, !=
	sentAt	DateTime	=, !=, >, >=, <, <=, BETWEEN
	createdAt	DateTime	=, !=, >, >=, <, <=, BETWEEN
	updatedAt	DateTime	N/A
	modifiedAt	DateTime	N/A
	redacted	Boolean	N/A
	createdByTeams	List	N/A
	rev	Integer	N/A
	reactions	List	N/A
	intentDetections	List	N/A
Notes	id	String	N/A
	body	String	CONTAINS
	createdAt	DateTime	=, !=, >, >=, <, <=, BETWEEN

Entity	Field	Data type	Supported operators
	updatedAt	DateTime	=, !=, >, >=, <, <=, BETWEEN
	modifiedAt	DateTime	=, !=, >, >=, <, <=, BETWEEN
	createdByTeams	List	N/A

Partitioning queries

Field-based partitioning

You can provide the additional Spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, and `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition the query.
- `LOWER_BOUND`: an **inclusive** lower bound value of the chosen partition field.

  For the DateTime field, we accept the value in ISO format.

  Example of valid value:

  ```
  "2023-01-15T11:18:39.205Z"
  ```
- `UPPER_BOUND`: an **exclusive** upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: the number of partitions.

Entity-wise partitioning field support details are captured in the following table:

Entity name	Partitioning fields	Data type
Companies	modifiedAt	DateTime

Entity name	Partitioning fields	Data type
Conversations	createdAt, updatedAt , modifiedAt, endedAt, lastMessageAt	DateTime
	messageCount, noteCount	BigInteger
	priority	Integer
Customers	createdAt, updatedAt, modifiedAt	DateTime
Messages	errorAt, sentAt, createdAt	DateTime
	size	BigInteger
Notes	createdAt, updatedAt, modifiedAt	DateTime

Example:

```
Kustomer_read = glueContext.create_dynamic_frame.from_options(
    connection_type="kustomer",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "conversation",
        "API_VERSION": "v1",
        "PARTITION_FIELD": "createdAt"
        "LOWER_BOUND": "2023-01-15T11:18:39.205Z"
        "UPPER_BOUND": "2023-02-15T11:18:39.205Z"
        "NUM_PARTITIONS": "2"
    }
```

Kustomer connection options

The following are connection options for Kustomer:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your object in Kustomer.
- `API_VERSION`(String) - (Required) Used for Read. Kustomer Rest API version you want to use.

- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for read. Number of partitions for read.
- `INSTANCE_URL`(String) - (Required) Used for Read. Kustomer instance URL.

Kustomer limitations

The following are limitations or notes for Kustomer:

- The `Customer Searches` entity is not supported since the Kustomer API documentation has not declared any endpoint for it.
- The support of filtration and incremental transfer on the `Klasses` entity is not supported.
- Order by can be supported on multiple applicable fields in a single request.

  However, the order by functionality on multiple fields has been observed to behave inconsistently from the SaaS end for some combinations. It is unpredictable as there could be 'n' combinations that might show incorrect sorting results. For example:

  For the `Customers` entity, order by `progressiveStatus desc, name asc` doesn't yield the correct sorted result. It sorts only based on the `progressiveStatus` order. If such a behaviour is observed, you can use a single field to be ordered upon.
- Order by on the field 'id' is only supported by the `Conversations` and `Messages` entities as a query parameter. For example: https://api.kustomerapp.com/v1/conversations?sort=desc (This sorts the results by 'id' in descending order.)

  Additionally, any other filter or ordering on any other field is translated into a POST request body having the API endpoint as POST https://api.kustomerapp.com/v1/customers/search To allow the support of ordering by 'id' in `Conversations` and `Messages`, either only order by id should be present or any other filter and/or order by on any other applicable field.

- Kustomer allows a maximum of 10K records to be fetched irrespective of a filtered or a non-filtered request. Due to this limitation, there will be data loss for any entity holding more than 10K records. There are two possible workarounds that you can perform to partially mitigate this:
  - Apply filters to fetch a specific set of records.
  - If there are more than 10K records with an applied filter, apply a successive filter value in a new subsequent request or apply ranges in filters. For example:

    1st request's filterExpression: `modifiedAt >= 2022-03-15T05:26:23.000Z and modifiedAt < 2023-03-15T05:26:23.000Z`

    Assume this exhausts the 10K record limit.

    Another request can be triggered with filterExpression: `modifiedAt >= 2023-03-15T05:26:23.000Z`
- As a SaaS behavior, the `CONTAINS` operator in Kustomer supports matching only on complete words and not partial matches within a word. For example: "body CONTAINS 'test record'" will match a record having 'test' in the 'body' field. However, "body CONTAINS 'test'" will not match a record having 'testAnotherRecord' in the 'body' field.

Connecting to LinkedIn

LinkedIn is a paid marketing tool that offers access to LinkedIn social networks through various sponsored posts and other methods. LinkedIn is a powerful marketing tool for B2B companies to build leads, online recognition, share content, and more.

Topics

- AWS Glue support for LinkedIn
- Policies containing the API operations for creating and using connections
- Configuring LinkedIn
- Configuring LinkedIn connections
- Reading from LinkedIn entities
- LinkedIn connection options
- Creating a LinkedIn account

- Limitations

AWS Glue support for LinkedIn

AWS Glue supports LinkedIn as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from LinkedIn.

Supported as a target?

No.

Supported LinkedIn API versions

202406 (June 2024)

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the preceding method, alternatively, use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring LinkedIn

Before you can use AWS Glue to transfer from LinkedIn, you must meet the following requirements:

Minimum requirements

- You have a LinkedIn account. For more information about creating an account, see Creating a LinkedIn account.
- Your LinkedIn account is enabled for API access.
- You have created a `OAuth2 API` integration in your LinkedIn account. This integration provides the client credentials that AWS Glue uses to access your data securely when it makes authenticated calls to your account. For more information, see the section called “Creating a LinkedIn account”.

If you meet these requirements, you’re ready to connect AWS Glue to your LinkedIn account. For typical connections, you don't need do anything else in LinkedIn.

Configuring LinkedIn connections

LinkedIn supports `AUTHORIZATION_CODE` grant type for OAuth2.

This grant type is considered “three-legged” `OAuth` as it relies on redirecting users to the third-party authorization server to authenticate the user. Users may opt to create their own connected app in LinkedIn and provide their own client ID and client secret when creating connections through the AWS Glue console. In this scenario, they will still be redirected to LinkedIn to login and authorize AWS Glue to access their resources.

This grant type results in both a refresh token and an access token. The access token expires 60 days after creation. A new access token can be obtained using the refresh token.

For public LinkedIn documentation on creating a connected app for `Authorization Code OAuth` flow, see Authorization Code Flow (3-legged OAuth).

Configuring a LinkedIn connection

1. In AWS Secrets Manager, create a secret with the following details:
   - For customer managed connected app – Secret should contain the connected app Consumer Secret with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as key.
   - For AWS Managed connected app – Empty secret or secret with some temporary value.

   > **Note**
   >
   > It is a must to create a secret per connection in AWS Glue.

2. In the AWS Glue Studio, create a connection under **Data Connections** by following the steps below:
   1. When selecting a **Connection type**, select **LinkedIn**.
   2. Provide the LinkedIn environment.
   3. Select the IAM role for which AWS Glue can assume and has permissions for following actions:

      JSON

      ```
      {
        "Version": "2012-10-17",
        "Statement": [
          {
            "Effect": "Allow",
            "Action": [
              "secretsmanager:DescribeSecret",
              "secretsmanager:GetSecretValue",
              "secretsmanager:PutSecretValue",
      ```

```
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

    4. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.
    5. Select the **Network options** if you want to use your network.
3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from LinkedIn entities

Prerequisites

A LinkedIn Object you would like to read from. Refer the supported entities table below to check the available entities.

Supported entities

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Ad Accounts	Yes	Yes	Yes	Yes	No
Campaigns	Yes	Yes	Yes	Yes	No
Campaign Groups	Yes	Yes	Yes	Yes	No
Creatives	Yes	Yes	Yes	Yes	No
Ad Analytics	Yes	No	No	Yes	No
Ad Analytics All AdAcocunts	Yes	No	No	Yes	No

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Ad Analytics All Campaigns	Yes	No	No	Yes	No
Ad Analytics All CampaignGroups	Yes	No	No	Yes	No
Ad Analytics All AdCreatives	Yes	No	No	Yes	No
Share Statistics	Yes	No	No	Yes	No
Page Statistics	Yes	No	No	Yes	No
Follower Statistics	Yes	No	No	Yes	No

Example

```
netsuiteerp_read = glueContext.create_dynamic_frame.from_options(
    connection_type="linkedin",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "adaccounts",
        "API_VERSION": "202406"
    }
)
```

LinkedIn entity and field details

Field Data Type	Supported Filter Operators
String	=
DateTime	BETWEEN, =
Numeric	=
Boolean	=

LinkedIn connection options

The following are connection options for LinkedIn:

- `ENTITY_NAME`(String) – (Required) Used for Read/Write. The name of your Object in LinkedIn. For example, adAccounts.
- `API_VERSION`(String) – (Required) Used for Read/Write. LinkedIn Rest API version you want to use. The value will be 202406, as LinkedIn currently supports only version 202406.
- `SELECTED_FIELDS`(List<String>) – Default: empty(SELECT *). Used for Read. Columns you want to select for the selected entity.
- `FILTER_PREDICATE`(String) – Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) – Default: empty. Used for Read. Full Spark SQL query.

Creating a LinkedIn account

Creating a LinkedIn App and OAuth credentials

1. Navigate to your **LinkedIn Developer Network** page and log in with your LinkedIn account credentials.
2. Navigate to the **My Apps** page and choose **Create Application** to create a new LinkedIn App.
3. Enter the following details into the app registration form:
    - **Company Name** – Select an existing company or create a new company.
    - **Name** – Enter the application name.

- **Description** – Enter the application description.
- **Application Logo** – Select an image file as your application logo.
- **Application Use** – Select the use of your application.
- **Website URL** – Enter the websiute URL that contains detailed information about your application.
- **Business Email** – Enter your business email address.
- **Business Phone** – Enter your business phone number.
- **LinkedIn API Terms of Use** – Read and agree.

4. Upon completion of the app registration form, choose **Submit**.

   You will be redirected to the **Authentication** page, where the Authentication Keys (Client ID and Client Secret) and other relevant details will be displayed.
5. If your web application requires access to the user's email address from their LinkedIn account, select the `r_emailaddress` permission. Additionally, you can specify Authorized Redirect URLs for your LinkedIn application.

Creating a page in LinkedIn account

1. Navigate to LinkedIn Developer Products.
2. At the upper-right corner of the **LinkedIn Developer Products** page, select **My apps**.
3. At the upper-right corner of the **My apps** page, select **Create app**.
4. On the **Create an app** page, enter your app name in the **App name** field.
5. In the **LinkedIn Page** field, enter your company page name or URL.

> **Note**
>
> If you don’t have a LinkedIn Page, you can create one by selecting **Create a new LinkedIn**.

6. In the **Privacy policy URL** field, enter your privacy policy URL.
7. Choose **Upload a logo** to upload an image that is to be displayed to users when they authorize with your app.
8. In the **Legal agreement** section, select **I have read and agree to these terms**.
9. Choose **Create app**.

Your new app will be created and will be available under the **My apps** tab.

Publishing campaign ads in LinkedIn

1. Log in to **Campaign Manager**.
2. Select an existing **Campaign Group**, or choose **Create** to create a new one.
3. Select your objective.
4. Select your group, budget, and schedule.
5. Build your target audience.
6. Select your ad format.
7. Select your budget and schedule.
8. Set up your ad(s).
9. Review and Launch.

Limitations

For the Analytics fields`ad_analytics_all_adAccounts`, `ad_analytics_all_campaigns`, `ad_analytics_all_campaign_groups`, and `ad_analytics_all_adCreatives` a filter is mandatory to retrieve the records.

Connecting to Mailchimp

Mailchimp is an all-in-one marketing platform that helps you manage and talk to your clients, customers, and other interested parties. Their approach to marketing focuses on healthy contact management practices, beautifully designed emails, unique automated workflows, and powerful data analysis. If you're a Mailchimp user, you can connect AWS Glue to your Mailchimp account. Then, you can use Mailchimp as a data source in your ETL jobs. Run these jobs to transfer data between Mailchimp and AWS services or other supported applications.

Topics

- AWS Glue support for Mailchimp
- Policies containing the API operations for creating and using connections
- Configuring Mailchimp
- Configuring Mailchimp connections

- Reading from Mailchimp entities
- Mailchimp connection options
- Creating an Mailchimp account
- Limitations

AWS Glue support for Mailchimp

AWS Glue supports Mailchimp as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Mailchimp.

Supported as a target?

No.

Supported Mailchimp API versions

3.0

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
        ],
        "Resource": "*"
      }
    ]
}
```

If you don't want to use the preceding method, alternatively, use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Mailchimp

Before you can use AWS Glue to transfer from Mailchimp, you must meet the following requirements:

Minimum requirements

- You have an Mailchimp account with email and password. For more information about creating an account, see Creating a Mailchimp account.
- You must have AWS Account created with the service access to AWS Glue.
- Ensure you have created one of the following resources. These resources provide credentials that AWS Glue uses to securely access your data when making authenticated calls to your account:
  - A Developer App that supports OAuth 2.0 authentication. For more information about creating a Developer App, see Creating a Mailchimp account.

If you meet these requirements, you’re ready to connect AWS Glue to your Mailchimp account. For typical connections, you don't need do anything else in Mailchimp.

Configuring Mailchimp connections

Mailchimp supports following two types for authentication mechanism:

- Mailchimp supports the `AUTHORIZATION_CODE` grant type.
  - This grant type is considered “three-legged” `OAuth` as it relies on redirecting users to the third party authorization server to authenticate the user. It is used when creating connections via the AWS Glue Console. The user creating a connection may by default rely on a AWS Glue owned connected app where they do not need to provide any `OAuth` related information except for their Mailchimp Client ID and Client Secret. The AWS Glue Console will redirect the user to Mailchimp where the user must login and allow AWS Glue the requested permissions to access their Mailchimp instance.
  - Users may still opt to create their own connected app in Mailchimp and provide their own Client ID and Client Secret when creating connections through the AWS Glue Console. In this scenario, they will still be redirected to Mailchimp to login and authorize AWS Glue to access their resources.
  - For public Mailchimp documentation on creating a connected app for `AUTHORIZATION_CODE` `OAuth` flow, see Access Data on Behalf of Other Users with OAuth 2 .
- **Custom Auth** – For public Mailchimp documentation about generating the required API keys for custom authorization, see About API Keys.

To configure a Mailchimp connection:

1. In AWS Secrets Manager, create a secret with the following details:
   - `OAuth` auth – For customer managed connected app: Secret should contain the connected app Consumer Secret with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as key.
   - Custom auth – For customer managed connected app: Secret should contain the connected app Consumer Secret with “api_key” as key.

   > **Note**
   >
   > It is a must to create a secret per connection in AWS Glue.

2. In AWS Glue Studio, create a connection under **Data Connections** by following the steps below:
   1. Under **Connections**, select **Create connection**.

   b. When selecting a **Data Source**, select Mailchimp.

   c. Provide the Mailchimp `instanceUrl`.

   d. Select the IAM role for which AWS Glue can assume and has permissions for following actions:

      JSON

      ```
      {
        "Version": "2012-10-17",
        "Statement": [
          {
            "Effect": "Allow",
            "Action": [
              "secretsmanager:DescribeSecret",
              "secretsmanager:GetSecretValue",
              "secretsmanager:PutSecretValue",
              "ec2:CreateNetworkInterface",
              "ec2:DescribeNetworkInterfaces",
              "ec2:DeleteNetworkInterface"
            ],
            "Resource": "*"
          }
        ]
      }
      ```

   e. Select Authentication Type to connect to Mailchimp:

      - For `OAuth` auth – Provide the Token URL, User Managed Client Application ClientId of the Mailchimp that you want to connect to.
      - For Custom auth – Select Authentication Type CUSTOM to connect to Mailchimp.

   f. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.

   g. Select the network options if you want to use your network.

3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

4. In your AWS Glue job configuration, provide `connectionName` as an Additional network connection.

Reading from Mailchimp entities

Prerequisites

A Mailchimp Object you would like to read from. Refer the supported entities table below to check the available entities.

Supported entities

- Abuse-reports
- Automation
- Campaigns
- Click-details
- Lists
- Members
- Open-details
- Segments
- Stores
- Unsubscribed

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Automation	Yes	Yes	Yes	Yes	Yes
Campaigns	No	No	No	No	No
Lists	Yes	Yes	No	Yes	Yes
Reports Abuse	No	Yes	No	Yes	Yes
Reports Open	No	Yes	No	Yes	Yes
Reports Click	Yes	Yes	No	Yes	Yes
Reports Unsubscribe	No	Yes	No	Yes	Yes
Segment	No	Yes	No	Yes	Yes

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Segment Members	Yes	Yes	No	Yes	No
Stores	Yes	Yes	Yes	Yes	No

Example

```
mailchimp_read = glueContext.create_dynamic_frame.from_options(
            connection_type="mailchimp",
            connection_options={

                  "connectionName": "connectionName",

                  "ENTITY_NAME": "stores",
"INSTANCE_URL": "https://us14.api.mailchimp.com",
                  "API_VERSION": "3.0"
               })
```

Mailchimp entity and field details

- Abuse-reports
- Automation
- Campaigns
- Click-details
- Lists
- Members
- Open-details
- Segments
- Stores
- Unsubscribed

Partitioning queries

You can provide the additional Spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, and `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition the query.
- `LOWER_BOUND`: an **inclusive** lower bound value of the chosen partition field.

  For the DateTime field, we accept the value in ISO format.

  Example of valid value:

  ```
  "2024-07-01T00:00:00.000Z"
  ```
- `UPPER_BOUND`: an **exclusive** upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: the number of partitions.

The following table describes the entity partitioning field support details:

Entity name	Partitioning fields	Data type

Example:

```
read_read = glueContext.create_dynamic_frame.from_options(
    connection_type="mailchimp",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "automations",
        "API_VERSION": "3.0",
        "INSTANCE_URL": "https://us14.api.mailchimp.com",
        "PARTITION_FIELD": "create_time",
```

```
        "LOWER_BOUND": "2024-02-05T14:09:30.115Z",
        "UPPER_BOUND": "2024-06-07T13:30:00.134Z",
        "NUM_PARTITIONS": "3"
    }
```

Mailchimp connection options

The following are connection options for Mailchimp:

- `ENTITY_NAME`(String) – (Required) Used for Read/Write. The name of your Object in Mailchimp.
- `INSTANCE_URL`(String) - (Required) A valid Mailchimp Instance URL.
- `API_VERSION`(String) - (Required) Used for Read. Mailchimp Engage Rest API version you want to use. For example: 3.0.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.

Creating an Mailchimp account

1. Navigate to the Mailchimp login page, enter your email ID and password, and then choose **Sign up**.
2. Open the confirmation email from Mailchimp, and choose the confirmation link to verify your account.

   > **Note**
   >
   > The time it takes to receive the activation email may vary. If you haven't received the activation email, check your spam folder, and read our activation email

troubleshooting tips. Mailchimp blocks signups from role-based email addresses such as, admin@pottedplanter.com or security@example.com.

The first time you log in to your account, Mailchimp asks for required information. Mailchimp uses this information to help ensure your account is compliant with their Terms of Use and to provide guidance that's relevant to you and your company's needs.

3. Enter your information, follow the prompts to finish the activation process, and get started in your new Mailchimp account.

Registering an `OAuth2.0` application

1. Navigate to the Mailchimp login page, enter your email ID and password, and choose **Log in**.
2. Select the **User** icon in the upper-right corner, and then choose **Account and billing** from the dropdown menu.
3. Select **Extras** and choose **Registered apps** from the dropdown menu.
4. Locate and choose **Register An App**.
5. Enter the following details:
    - **App name** – Name of the app.
    - **Company / Organization** – Name of your Company or Organization.
    - **App website** – Website of the app.
    - **Redirect URI** – A Redirect URI pattern is a URI path (or comma-separated list of paths) to which Mailchimp can redirect (if requested) when the login flow is complete. For example, `https://ap-southeast-2\\.console\\.aws\\.amazon\\.com`
6. Choose **Create**.
7. The **Client ID** and **Client Secret** will now be visible. Copy and save them in a secure location. Then, choose **Done**.

> **Note**
>
> Your Client ID and Client Secret strings are credentials used to establish a connection with this connector when using AppFlow or AWS Glue.

Generating an API key

1. Navigate to the Mailchimp login page, enter your email ID and password, and choose **Log in**.
2. Select the **User** icon in the upper-right corner, and then choose **Account and billing** from the dropdown menu.
3. Select **Extras** and choose **API keys** from the dropdown menu.
4. Choose **Create A Key**.
5. Enter a name for the key and choose **Generate Key**.

   The next page displays the generated API key.
6. Copy your key, store it securely, and choose **Done**.

Limitations

The following are limitations for the Mailchimp connector:

- Filtration is only supported by `Campaigns`, `Automations`, `Lists`, `Open Details`, `Members`, and `Segments` entities.
- While using a filter on `DateTime` datatype field, you must pass values in this format: `yyyy-mm-ddThh:MM:ssZ`

Connecting to Microsoft Dynamics 365 CRM

Microsoft Dynamics 365 is a product line of enterprise resource planning and customer relationship management intelligent business applications.

Topics

- AWS Glue support for Microsoft Dynamics 365
- Policies containing the API operations for creating and using connections
- Configuring Microsoft Dynamics 365 CRM
- Configuring Microsoft Dynamics 365 CRM connections
- Reading from Microsoft Dynamics 365 CRM entities
- Microsoft Dynamics 365 CRM connection option reference
- Limitations

AWS Glue support for Microsoft Dynamics 365

AWS Glue supports Microsoft Dynamics 365 as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Microsoft Dynamics 365.

Supported as a target?

No.

Supported Microsoft Dynamics 365 CRM API versions

v9.2.

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

You can also use the following managed IAM policies to allow access:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Microsoft Dynamics 365 CRM

Before you can use AWS Glue to transfer data from Microsoft Dynamics 365 CRM, you must meet these requirements:

Minimum requirements

- You have a Microsoft Dynamics 365 CRM developer account with ClientId and Secret.
- Your Microsoft Dynamics 365 CRM account has API access with a valid license.

If you meet these requirements, you’re ready to connect AWS Glue to your Microsoft Dynamics 365 CRM account. For typical connections, you don't need do anything else in Microsoft Dynamics 365 CRM.

Configuring Microsoft Dynamics 365 CRM connections

AUTHORIZATION_CODE Grant Type

- This grant type is considered “three-legged” OAuth as it relies on redirecting users to the third party authorization server to authenticate the user. It is used when creating connections via the AWS Glue Console. The AWS Glue Console will redirect the user to Microsoft Dynamics 365 CRM where the user must login and allow AWS Glue the requested permissions to access their Microsoft Dynamics 365 CRM instance.
- Users may opt to create their own connected app in Microsoft Dynamics 365 CRM and provide their own client id and client secret when creating connections through the AWS Glue Console. In this scenario, they will still be redirected to Microsoft Dynamics 365 CRM to login and authorize AWS Glue to access their resources.

- This grant type results in a refresh token and access token. The access token is short lived, and may be refreshed automatically without user interaction using the refresh token.
- For public Microsoft Dynamics 365 CRM documentation on creating a connected app for Authorization Code OAuth flow, see | Microsoft Learn. Microsoft App Registration.

Microsoft Dynamics 365 CRM supports OAuth2.0 authentication.

To configure a Microsoft Dynamics 365 CRM connection:

1. In AWS Secrets Manager, create a secret with the following details. It is required to create a secret for each connection in AWS Glue.
   - For AuthorizationCode grant type:

     For customer managed connected app - Secret should contain the connected app Client Secret with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as key.
2. In AWS Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Data Source**, select Microsoft Dynamics 365 CRM.
   b. Select the **INSTANCE_URL** of the Microsoft Dynamics 365 CRM instance you want to connect to.
   c. Select the IAM role that AWS Glue can assume and has permissions for following actions:

      JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

   d. Select **Token URL** and **Authorization Code URL** to access your Microsoft Dynamics 365 CRM workspace.
   e. Provide **User Managed Client Application ClientId** of your Microsoft Dynamics 365 CRM app.
   f. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.
   g. Select the network options if you want to use your network.
3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`. Choose **Next**.
4. In your AWS Glue job configuration, provide `connectionName` as an **Additional network connection**.

Reading from Microsoft Dynamics 365 CRM entities

Prerequisites

- A Microsoft Dynamics 365 CRM object you would like to read from. You will need the object name such as contacts or accounts. The following table shows the supported entities.

Supported entities

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Dynamic entity	Yes	Yes	Yes	Yes	Yes

Example

```
dynamics365_read = glueContext.create_dynamic_frame.from_options(
    connection_type="microsoftdynamics365crm",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "dynamic_entity",
        "API_VERSION": "v9.2",
        "INSTANCE_URL": "https://{tenantID}.api.crm.dynamics.com"
    }
```

Microsoft Dynamics 365 CRM Entity and Field Details

Entities with dynamic metadata:

Microsoft Dynamics 365 CRM provides endpoints to fetch metadata dynamically. Therefore, for dynamic entities, the operator support is captured at datatype level.

Entity	Data Type	Supported Operators
Dynamic entity	DateTime	=, <, <=, >, >=, BETWEEN
	Date	=, <, <=, >, >=
	String	=, !=
	Double	=, <, <=, >, >=
	Integer	=, <, <=, >, >=
	Decimal	=, <, <=, >, >=
	Long	=, <, <=, >, >=
	BigInteger	=, <, <=, >, >=
	List	NA
	Struct	NA
	Map	NA

Partitioning queries

Microsoft Dynamics 365 CRM supports only field based partitioning.

Additional spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, `NUM_PARTITIONS` can be provided if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition query.
- `LOWER_BOUND`: an inclusive lower bound value of the chosen partition field.

  For Datetime, we accept the Spark timestamp format used in Spark SQL queries. Example of valid values: `"2024-01-30T06:47:51.000Z"`.
- `UPPER_BOUND`: an exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: number of partitions.

Entity wise partitioning field support details are captured in below table:

Entity Name	Partitioning Fields	DataType
Dynamic Entity (Standard entity)	Dynamic DateTime fields which are queryable	createdon, modifiedon
Dynamic Entity (Custom entity)	createdon, modifiedon	createdon, modifiedon

Example

```
dynamics365_read = glueContext.create_dynamic_frame.from_options(
    connection_type="microsoftdynamics365crm",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "dynamic_entity",
        "API_VERSION": "v9.2",
        "instanceUrl": "https://{tenantID}.api.crm.dynamics.com"
        "PARTITION_FIELD": "createdon"
        "LOWER_BOUND": "2024-01-30T06:47:51.000Z"
        "UPPER_BOUND": "2024-06-30T06:47:51.000Z"
        "NUM_PARTITIONS": "10"
    }
```

Microsoft Dynamics 365 CRM connection option reference

The following are connection options for Microsoft Dynamics 365 CRM:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your object in Microsoft Dynamics 365 CRM.
- `API_VERSION`(String) - (Required) Used for Read. Microsoft Dynamics 365 CRM Rest API version you want to use.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `INSTANCE_URL`(String) - (Required) A valid Microsoft Dynamics 365 CRM Instance URL with the format: `https://{tenantID}.api.crm.dynamics.com`
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field. Example: `2024-01-30T06:47:51.000Z`.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field. Example: `2024-06-30T06:47:51.000Z`.

Limitations

The following are limitations for the Microsoft Dynamics 365 CRM connector:

- In Microsoft Dynamics 365 CRM, record based partitioning is not supported as it does not support an offset parameter, and thus record based partitioning cannot be supported.
- Pagination is set at a maximum of 500 records per page to avoid Internal Server exceptions from SaaS due to a combination of data size and rate limitations.
  - SaaS documentation on pagination
  - SaaS documentation on rate limits
- Microsoft Dynamics 365 CRM supports `order by` on only parent fields for all entities. `order by` is not supported on sub-fields.
  - Both ASC and DESC directions are supported.
  - `order by` on multiple fields is supported.

- Filtering on "createddatetime" field of the `aadusers` standard entity is throwing a bad request error from SaaS even though it supports filtration. There is no specific identification of any other entity having a similar issue due to the dynamic nature of the metadata and neither is the root cause known. Hence, it cannot be handled.
- Complex object types, such as Struct, List, and Map do not support filtration.
- Many fields that can be retrieved from a response have `isRetrievable` marked as `false` in dynamic metadata response. In order to avoid data loss, `isRetrievable` is set to `true` for all fields.
- Field based partitioning will be supported on all entities when it adheres to the following criteria:
    - DateTime queryable fields should be present in standard entities or `createdon` and `modifiedon` fields (system-generated) in custom entities.
    - There is no exclusive identification of system generated fields or the nullable property from any SaaS metadata APIs, however it is a general practice that only the fields available by default, are filterable and non-nullable. Therefore, the above criterion of field selection is considered null safe and if it is filterable, it will be eligible for partitioning.

Connecting to Microsoft Teams

Microsoft Teams is a collaborative workspace within Microsoft 365 that acts as a central hub for workplace conversations, collaborative teamwork, video chats and document sharing, all designed to aid worker productivity in a unified suite of tools.

Topics

- AWS Glue support for Microsoft Teams
- Policies containing the API operations for creating and using connections
- Configuring Microsoft Teams
- Configuring Microsoft Teams connections
- Reading from Microsoft Teams entities
- Microsoft Teams connection option reference
- Limitations
- Create a new Microsoft Teams account:

AWS Glue support for Microsoft Teams

AWS Glue supports Microsoft Teams as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Microsoft Teams.

Supported as a target?

No.

Supported Microsoft Teams API versions

v1. For entity support per version specific, see Supported entities for source.

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

You can also use the following managed IAM policies to allow access:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Microsoft Teams

Before you can use AWS Glue to transfer data from Microsoft Teams, you must meet these requirements:

Minimum requirements

- You have a Microsoft Teams developer account with Email and Password. For more information, see Create a new Microsoft Teams account:.
- You should have setup an OAuth2 app in your Microsoft account that provides the client ID and secret credentials that AWS Glue uses to access your data securely when it makes authenticated calls to your account. For more information, see Create a new Microsoft Teams account:.

If you meet these requirements, you're ready to connect AWS Glue to your Microsoft Teams account. For typical connections, you don't need do anything else in Microsoft Teams.

Configuring Microsoft Teams connections

Microsoft Teams supports following two types for authentication mechanism:

1. OAuth Auth: Microsoft Teams supports AUTHORIZATION_CODE grant type for OAuth2.
    - This grant type is considered “three-legged” OAuth as it relies on redirecting users to the third party authorization server to authenticate the user. It is used when creating connections via the AWS Glue Console. The user creating a connection may by default rely on a AWS Glue-owned connected app where they do not need to provide any OAuth related information except for the Microsoft Teams instanceurl. The AWS Glue Console will redirect the user to

Microsoft Teams where the user must login and allow AWS Glue the requested permissions to access their Microsoft Teams instance.

- Users may opt to create their own connected app in Microsoft Teams and provide their own client id and client secret when creating connections through the AWS Glue Console. In this scenario, they will still be redirected to Microsoft Teams to login and authorize AWS Glue to access their resources.
- This grant type results in a refresh token and access token. The access token is active for one hour, and may be refreshed automatically without user interaction using the refresh token.
- For public Microsoft Teams documentation on creating a connected app for Authorization Code OAuth flow, see | Microsoft Learn. Register an application with the Microsoft identity platform - Microsoft Graph.

To configure a Microsoft Teams connection:

1. In AWS Secrets Manager, create a secret with the following details. It is required to create a secret for each connection in AWS Glue.
   a. For OAuth auth:
      - For customer managed connected app - Secret should contain the connected app Consumer Secret with USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET as key.
2. In AWS Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. Under Data Connections, choose **Create connection**.
   b. When selecting a **Data Source**, select Microsoft Teams.
   c. Provide your Microsoft Teams **Tenant ID**.
   d. Select the IAM role that AWS Glue can assume and has permissions for following actions:

      JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
            "ec2:DescribeNetworkInterfaces",
            "ec2:DeleteNetworkInterface"
          ],
          "Resource": "*"
        }
      ]
    }
```

   e. Provide User Managed Client Application ClientId of Microsoft Teams app.

   f. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.

   g. Select the network options if you want to use your network.

3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`. Choose **Next**.

4. In your AWS Glue job configuration, provide `connectionName` as an **Additional network connection**.

Reading from Microsoft Teams entities

Prerequisites

- A Microsoft Teams object you would like to read from. You will need the object name such as team or channel-message. The following table shows the supported entities.

Supported entities for Source

All entities are supported with API version 1.0.

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Teams	No	No	No	Yes	No
Team Members	Yes	Yes	No	Yes	Yes
Groups	Yes	Yes	Yes	Yes	Yes

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Group Members	Yes	Yes	No	Yes	No
Channels	Yes	No	No	Yes	Yes
Channel Messages	No	Yes	No	Yes	No
Channel Message Replies	No	Yes	No	Yes	No
Channel Tabs	Yes	No	No	Yes	No
Chats	Yes	Yes	Yes	Yes	Yes
Calendar Events	Yes	Yes	Yes	Yes	Yes

Example

```
MicrosoftTeams_read = glueContext.create_dynamic_frame.from_options(
    connection_type="MicrosoftTeams",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "company",
        "API_VERSION": "v1.0"
    }
```

Microsoft Teams Entity and Field Details

Entities list:

- Team: https://docs.microsoft.com/en-us/graph/api/user-list-joinedteams?view=graph-rest-1.0
- Team-Member: https://docs.microsoft.com/en-us/graph/api/team-list-members?view=graph-rest-1.0

- Group: https://docs.microsoft.com/en-us/graph/api/group-list?view=graph-rest-1.0
- Group-Member: https://docs.microsoft.com/en-us/graph/api/group-list-members?view=graph-rest-1.0
- Channel: https://docs.microsoft.com/en-us/graph/api/channel-list?view=graph-rest-1.0
- Channel-Message: https://docs.microsoft.com/en-us/graph/api/channel-list-messages?view=graph-rest-1.0
- Channel-Message-Reply: https://docs.microsoft.com/en-us/graph/api/chatmessage-list-replies?view=graph-rest-1.0
- Channel-Tab: https://docs.microsoft.com/en-us/graph/api/channel-list-tabs?view=graph-rest-1.0
- Chat: https://docs.microsoft.com/en-us/graph/api/chat-list?view=graph-rest-1.0
- Calendar-Event: https://docs.microsoft.com/en-us/graph/api/group-list-events?view=graph-rest-1.0

Partitioning queries

Additional spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, `NUM_PARTITIONS` can be provided if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition query.
- `LOWER_BOUND`: an inclusive lower bound value of the chosen partition field.

  For date, we accept the Spark date format used in Spark SQL queries. Example of valid values: `"2024-02-06"`.
- `UPPER_BOUND`: an exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: number of partitions.

Entity wise partitioning field support details are captured in below table:

Entity Name	Partitioning Fields	Data Type
Team Members	visibleHistoryStartDateTime	DateTime

Entity Name	Partitioning Fields	Data Type
Groups	createdDateTime	DateTime
Channels	createdDateTime	DateTime
Chats	createdDateTime, lastModifiedDateTime	DateTime
Calendar Events	createdDateTime, lastModifiedDateTime, originalStart	DateTime

Example

```
microsoftteams_read = glueContext.create_dynamic_frame.from_options(
    connection_type="MicrosoftTeams",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "group",
        "API_VERSION": "v1.0",
        "PARTITION_FIELD": "createdDateTime"
        "LOWER_BOUND": "2022-07-13T07:55:27.065Z"
        "UPPER_BOUND": "2022-08-12T07:55:27.065Z"
        "NUM_PARTITIONS": "2"
    }
```

Microsoft Teams connection option reference

The following are connection options for Microsoft Teams:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your Object in Microsoft Teams.
- `API_VERSION`(String) - (Required) Used for Read. Microsoft Teams Rest API version you want to use. Example: v1.0.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.

- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.

Limitations

The following are limitations for the Microsoft Teams connector:

- The Microsoft Teams API returns less number of records than specified for Chat and Team Member entities. This issue is reported to Microsoft Teams Support and it is under investigation.

Create a new Microsoft Teams account:

1. Navigate to Microsoft Teams's homepage, https://account.microsoft.com/account/ and choose **Sign in**.
2. Choose **Create one!**.
3. Enter the required information for account creation and create a new account.
4. Navigate to the Microsoft Teams website at https://www.microsoft.com/en-in/microsoft-teams/log-in.
5. Sign up using the Microsoft Account you just created.
6. After successful sign up on Teams, navigate to https://account.microsft.com/services.
7. Choose **Try Microsoft 365**.
8. Activate one of below Microsoft 365 or Microsoft Teams subscription to access all required features of Microsoft Teams connector:
   - Microsoft Teams Essentials
   - Microsoft 365 Business
   - Microsoft 365 Business Basic
   - Microsoft 365 Business Standard
   - Microsoft 365 Business Premium

Create a managed client app:

1. To create a managed application, you need to register a new OAuth app on Microsoft Entra (formerly Azure Active Directory):
2. Sign in to the Microsoft Entra admin center.
3. If you have access to multiple tenants, use the Settings icon in the top menu to switch to the tenant in which you want to register the application from the Directories + subscriptions menu.
4. Navigate to Identity > Applications > App registrations and select **New registration**.
5. Enter a display Name for your application.
6. Specify who can use the application in the Supported account types section. To make this app global select “Accounts in any organizational directory” or “Accounts in any organizational directory and personal Microsoft accounts”.
7. Enter Redirect URI `https://{region}.console.aws.amazon.com/appflow/oauth`. For example, for the `us-west-2 region`, add `https://us-west-2.console.aws.amazon.com/appflow/oauth`. You can add multiple URLs for different regions that you want to use.
8. Register the app.
9. Note the Client ID for future use.
10. Choose **Add a certificate or secret** in the Essentials section.
11. Choose **New Client Secret**.
12. Enter Description and Expires duration.
13. Copy and save the client secret for future use.
14. In the left side menu list, select **API permissions**.
15. Choose **Add a permission**.
16. Select “Microsoft Graph“.
17. Select “Delegated permissions”.
18. Check all the following permissions:

    - User.Read
    - Offline_access
    - User.Read.All
    - User.ReadWrite.All
    - TeamsTab.ReadWriteForTeam

- TeamsTab.ReadWriteForChat
- TeamsTab.ReadWrite.All
- TeamsTab.Read.All
- TeamSettings.ReadWrite.All
- TeamSettings.Read.All
- TeamMember.ReadWrite.All
- TeamMember.Read.All
- Team.ReadBasic.All
- GroupMember.ReadWrite.All
- GroupMember.Read.All
- Group.ReadWrite.All
- Group.Read.All
- Directory.ReadWrite.All
- Directory.Read.All
- Directory.AccessAsUser.All
- Chat.ReadWrite
- Chat.ReadBasic
- Chat.Read
- ChannelSettings.ReadWrite.All
- ChannelSettings.Read.All
- ChannelMessage.Read.All
- Channel.ReadBasic.All

19. Choose **Add permissions**. Your app is now setup successfully. You can use the client ID and client secret to create a new connection. For more information, see https://learn.microsoft.com/en-us/graph/auth-register-app-v2.

Connecting to Mixpanel

Mixpanel is a powerful real-time analytics platform that helps companies measure and optimize user engagement. Mixpanel is an app used for tracking customer behavior. It enables you to track how users engage with your product and analyze this data with interactive reports that let you query and visualize the results with just a few clicks. As a Mixpanel user, you can connect AWS Glue to your Mixpanel account. Then, you can use Mixpanel as a data source in your ETL jobs. Run these jobs to transfer data between Mixpanel and AWS services or other supported applications.

Topics

- AWS Glue support for Mixpanel
- Policies containing the API operations for creating and using connections
- Configuring Mixpanel
- Configuring Mixpanel connections
- Reading from Mixpanel entities
- Mixpanel connection options
- Creating a Mixpanel account and configuring the client app
- Limitations

AWS Glue support for Mixpanel

AWS Glue supports Mixpanel as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Mixpanel.

Supported as a target?

No.

Supported Mixpanel API versions

2.0

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the preceding method, alternatively, use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Mixpanel

Before you can use AWS Glue to transfer from Mixpanel, you must meet these requirements:

Minimum requirements

- You have a Mixpanel account. For more information about creating an account, see Creating a Mixpanel account.
- Your Mixpanel account is enabled for API access. API access is enabled by default for the Enterprise, Unlimited, Developer, and Performance editions.

If you meet these requirements, you're ready to connect AWS Glue to your Mixpanel account. For typical connections, you don't need do anything else in Mixpanel.

Configuring Mixpanel connections

Mixpanel supports username and password for `BasicAuth`. Basic Authentication is a simple authentication method where clients provide credentials directly to access protected resources. AWS Glue is able to use the username and password to authenticate Mixpanel APIs.

For public Mixpanel documentation about `BasicAuth` flow, see Mixpanel Service Accounts.

To configure a Mixpanel connection:

1. In AWS Secrets Manager, create a secret with the following details:
   - For Basic Authentication, Secret should contain the connected app Consumer Secret with `USERNAME` and `PASSWORD` as key.

     > **Note**
     >
     > It is a must to create a secret per connection in AWS Glue.
2. In the AWS Glue Studio, create a connection under **Data Connections** by following the steps below:
   1. When selecting a **Connection type**, select **Mixpanel**.
   2. Provide the `INSTANCE_URL` of the Mixpanel that you want to connect to.
   3. Select the IAM role for which AWS Glue can assume and has permissions for the following actions:

      JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
          "ec2:DescribeNetworkInterfaces",
          "ec2:DeleteNetworkInterface"
        ],
        "Resource": "*"
      }
    ]
  }
```

d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.

e. Select **Network options** if you want to use your network.

3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Mixpanel entities

Prerequisites

You must have a Mixpanel object, such as `Funnels`, `Retention`, or `Retention Funnels`, from which you would like to read data. Additionally, you will need to know the object name.

Supported entities

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Funnels	Yes	No	No	Yes	No
Retention	Yes	No	No	Yes	No
Segmentation	Yes	No	No	Yes	No
Segmentation Sum	Yes	No	No	Yes	No
Segmentation Average	Yes	No	No	Yes	No
Cohorts	Yes	No	No	Yes	No

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Engage	No	Yes	No	Yes	No
Events	Yes	No	No	Yes	No
Events Top	Yes	No	No	Yes	No
Events Names	Yes	No	No	Yes	No
Events Properties	Yes	No	No	Yes	No
Events Properties Top	Yes	No	No	Yes	No
Events Properties Values	Yes	No	No	Yes	No
Annotations	Yes	No	No	Yes	No
Profile Event Activity	Yes	No	No	Yes	No

Example

```
mixpanel_read = glueContext.create_dynamic_frame.from_options(
    connection_type="mixpanel",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "/cohorts/list?project_id=2603353",
        "API_VERSION": "2.0",
        "INSTANCE_URL": "https://www.mixpanel.com/api/app/me"
    }
```

Mixpanel entity and field details

Entity	Field	Data Type	Supported Operators
Funnel	funnel_id	Integer	'='
	workspace_id	Integer	'='
	from_date	Date	'='
	to_date	Date	'='
	length	Integer	'='
	length_unit	String	'='
	interval	Integer	'='
	unit	String	'='
	limit	Integer	'='
	data	Struct	
	meta	Struct	
Retention	workspace_id	Integer	'='
	unit	String	'='
	addiction_unit	String	'='
	from_date	Date	'='
	to_date	Date	'='
	event	String	'='
	limit	Integer	'='
	data	Struct	
Segmentation	workspace_id	Integer	'='

Entity	Field	Data Type	Supported Operators
	event	String	'='
	from_date	Date	'='
	to_date	Date	'='
	unit	String	'='
	interval	Integer	'='
	limit	Integer	'='
	type	String	'='
	series	List	
	values	Struct	
	data	Struct	
Segmentation Numeric	workspace_id	Integer	'='
	event	String	'='
	on	String	'='
	from_date	Date	'='
	to_date	Date	'='
	unit	String	'='
	type	String	'='
	series	List	
	values	Struct	
Segmentation Sum	workspace_id	Integer	'='

Entity	Field	Data Type	Supported Operators
	event	String	'='
	on	String	'='
	from_date	Date	'='
	to_date	Date	'='
	unit	String	'='
	metadata	Struct	
	results	Struct	
Segmentation Average	workspace_id	Integer	'='
	event	String	'='
	on	String	'='
	from_date	Date	'='
	to_date	Date	'='
	unit	String	'='
	metadata	Struct	
	results	Struct	
Cohorts	count	Integer	
	is_visible	Integer	
	description	String	
	created	DateTime	
	project_id	Integer	

Entity	Field	Data Type	Supported Operators
	id	BigInteger	
	name	String	
	data_group_id	String	
Engage	distinct_id	String	
		properties	Struct
Event	workspace	Integer	'='
	event	String	'='
	type	String	'='
	unit	String	'='
	interval	Integer	'='
	from_date	Date	'='
	to_date	Date	'='
	series	List	
	values	Struct	
Events Top	type	String	'='
	workspace_id	Integer	'='
	limit	Integer	'='
	amount	Integer	
	event	String	
	percent_change	Float	

Entity	Field	Data Type	Supported Operators
Event Name	data	List	
	workspace_id	Integer	'='
	type	String	'='
	limit	Integer	'='
Event Properties	workspace_id	Integer	'='
	event	String	'='
	name	String	'='
	type	String	'='
	unit	String	'='
	interval	Integer	'='
	from_date	Date	'='
	to_date	Date	'='
	limit	Integer	'='
	data	Struct	
	series	List	
	values	Struct	
Event Properties Top	workspace_id	Integer	'='
	event	String	'='
	limit	Integer	'='
	data	Struct	

Entity	Field	Data Type	Supported Operators
Event Properties Value	workspace_id	Integer	'='
	event	String	'='
	limit	Integer	'='
	name	String	'='
	data	List	
Annotation	workspace_id	Integer	
	date	DateTime	
	project_id	Integer	
	id	BigInteger	
	description	String	
	from_date	Date	BETWEEN
Profile Event Activity	workspace_id	Integer	'='
	distinct_ids	String	'='
	from_date	Date	'='
	to_date	Date	'='
	event	String	
	properties	Struct	

Mixpanel connection options

The following are connection options for Mixpanel:

- `ENTITY_NAME` (String) – (Required) Used for Read/Write. The name of your Object in Mixpanel.
- `API_VERSION` (String) – (Required) Used for Read/Write. Mixpanel Rest API version you want to use. For example: v2.0.
- `SELECTED_FIELDS`(List<String>) – Default: empty (SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) – Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) – Default: empty. Used for Read. Full Spark SQL query.

Creating a Mixpanel account and configuring the client app

Creating a Mixpanel account

1. Navigate to the [Mixpanel home page/](#).
2. On the **Mixpanel** home page, choose **Sign Up** at the upper-right corner of the page.
3. On the **Let's get started** page, complete the following actions:
    - Enter your email address in the designated field.
    - Select the required checkbox to agree to the terms.
    - Choose **Get Started** to proceed.

    Upon successful completion, you will receive a verification email.
4. Check your email inbox for a verification message, open the email, and follow the instructions to verify your email address.
5. On the verification page, choose **Verify Email** to complete your email verification.
6. On the **Name Your Organization** page, enter your organization name and choose **Next**.
7. On the **Your First Project** page, enter your project details and choose **Create**.
8. On the next page, choose **Let's get Started** to complete the creation of your account.

Logging into a Mixpanel account

1. Navigate to the [Mixpanel login page/](#).
2. Enter your email address and choose **Continue**.

3. Check your email inbox for a verification message, open the email, and follow the instructions to verify your email address.
4. On the next page, choose **Log In button** to log in to your account.

Purchasing a Mixpanel plan

1. On the Mixpanel page, select the **Settings** icon located in the upper-right corner of the page.
2. From the list of options, select **Plan Details and Billing**.
3. On the **Plan Details and Billing** page, select **Upgrade or Modify**.
4. In the next page, select the plan that you want to purchase.

   This completes the account creation and plan purchasing process.

Creating a username and client secret (To register your app)

1. On the Mixpanel page, select the **Settings** icon located in the upper-right corner of the page."
2. From the list of options, select **Project Settings**.
3. On the **Project Settings** page, select **Service Accounts** and then select **Add Service Account**.
4. From the **Service Account** dropdown list, select the **service Account or enter name to create**, add **Project Role**, specify **expires**, and select **Add**.

   > ⚠ **Important**
   >
   > After completing the previous step, the following page displays the service account's secret key. Ensure to save the service account's secret key. You will not be able to access it again after this point.

Limitations

The following are limitations for the Mixpanel connector:

- For `Segmentation Numeric` entity, the Mixpanel API throws a `400 (Bad Request)` error if no numeric data is found for the mandatory filters. We are treating this as an OK response to prevent flow failure.
- The queryable field `limit` has been removed from the supported entities because:

  - It was causing errors due to being interpreted as the SDK's limit feature
  - The filter served no practical purpose
  - Equivalent functionality is now covered by the limit feature implementation
- Field-based partitioning cannot be supported due to the absence of required operators (>=, <=, <, >, `between`) for partitioning from the SaaS platform. Although it supports the `between` operator, the fields for which it supports this operator are non-retrievable. Hence, the criteria for field-based partitioning are not satisfied.
- As there is no provision to get an 'offset' value for entities that support pagination, it is not possible to support record-based partitioning for Mixpanel.
- `Cohorts` entity only supports `CreatedDate/Time` field and there is no field to identify `UpdatedDate/Time` as a result `DML_Status` cannot be identified. Also, there is no endpoint to identify deleted records. Hence, CDC cannot be supported.
- To run a AWS Glue job for the entities mentioned below, mandatory filters are required. Refer to the table below for entity names and their required filters.

Entity name and required filters

Entity name	Mandatory Filters
Annotations	from_date, to_date
Cohorts	None
Engage	None
Event	event, type, unit, from_date, to_date, on
Events Name	type
Events Properties	event, name, type, unit, from_date, to_date
Events Properties Top	event
Events Properties Values	event, name
Events Top	type
Funnels	funnel_id, from_date, to_date

Entity name	Mandatory Filters
Profile Event Activity	distinct_ids, from_date, to_date
Retention	from_date, to_date, unit, addiction_unit
Segmentation	event, from_date, to_date
Segmentation Average	event, from_date, to_date, on
Segmentation Numeric	event, from_date, to_date, on
Segmentation Sum	event, from_date, to_date, on

Connecting to Monday

Monday.com is a versatile work operating system that streamlines project management and team collaboration. It features customizable workflows, visual dashboards, and automation tools to enhance productivity. Users can track tasks, manage resources, and communicate effectively in one integrated platform.

Topics

- AWS Glue support for Monday
- Policies containing the API operations for creating and using connections
- Configuring Monday
- Configuring Monday connections
- Reading from Monday entities
- Monday connection option reference
- Limitations
- Create a new Monday account:

AWS Glue support for Monday

AWS Glue supports Monday as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Monday.

Supported as a target?

No.

Supported Monday API versions

v2.

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

You can also use the following managed IAM policies to allow access:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.

- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Monday

Before you can use AWS Glue to transfer data from Monday, you must meet these requirements:

Minimum requirements

- You have a Monday developer account with Email and Password. For more information, see Create a new Monday account:.
- Your Monday developer account is enabled for API access. All use of the Monday APIs available at no additional cost within trial period. Once trial period is over you need to buy subscription to create and access data. For more information, see Monday's licensing page for more details.

If you meet these requirements, you're ready to connect AWS Glue to your Monday account. For typical connections, you don't need do anything else in Monday.

Configuring Monday connections

Monday supports following two types for authentication mechanism:

1. OAuth Auth: Monday supports AUTHORIZATION_CODE grant type for OAuth2.
    - This grant type is considered "three-legged" OAuth as it relies on redirecting users to the third party authorization server to authenticate the user. It is used when creating connections via the AWS Glue Console. The user creating a connection may by default rely on a AWS Glue-owned connected app where they do not need to provide any OAuth related information except for the Monday instanceurl. The AWS Glue Console will redirect the user to Monday where the user must login and allow AWS Glue the requested permissions to access their Monday instance.
    - Users should opt to create their own connected app in Monday and provide their own client id and client secret when creating connections through the AWS Glue Console. In this scenario, they will still be redirected to Monday to login and authorize AWS Glue to access their resources.

- This grant type results in a refresh token and access token. The access token is active for one hour, and may be refreshed automatically without user interaction using the refresh token.
- For more information, see documentation on creating a connected app for AUTHORIZATION_CODE OAuth flow.

2. Custom Auth:
   - For public Monday documentation on generating the required API keys for custom authorization, see https://developer.monday.com/api-reference/docs/authentication#api-token-permissions .

To configure a Monday connection:

1. In AWS Secrets Manager, create a secret with the following details. It is required to create a secret for each connection in AWS Glue.
   a. For OAuth auth:
      - For customer managed connected app - Secret should contain the connected app Consumer Secret with USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET as key.

   b. For Custom auth:
      - For customer managed connected app - Secret should contain the connected app Consumer Secret with `personalAccessToken` as key.
2. In AWS Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. Under Data Connections, choose **Create connection**.

   b. When selecting a **Data Source**, select Monday.

   c. Provide your Monday **instanceURL**.

   d. Select the IAM role that AWS Glue can assume and has permissions for following actions:

   JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
                "ec2:CreateNetworkInterface",
                "ec2:DescribeNetworkInterfaces",
                "ec2:DeleteNetworkInterface"
            ],
            "Resource": "*"
        }
    ]
}
```

   e. Select Authentication Type to connect to Monday
      - For OAuth auth: Provide the **Token URL** and **User Managed Client Application ClientId** of the Monday you want to connect to.
      - For Custom auth: Select Authentication Type **CUSTOM** to connect to Monday.
   f. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.
   g. Select the network options if you want to use your network.
3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`. Choose **Next**.
4. In your AWS Glue job configuration, provide `connectionName` as an **Additional network connection**.

Reading from Monday entities

Prerequisites

- A Monday Object you would like to read from. Refer the supported entities table below to check the available entities.

Supported entities for Source

Entities list:

- Account: https://developer.monday.com/api-reference/docs/account#queries
- Board: https://developer.monday.com/api-reference/docs/boards#queries
- Column: https://developer.monday.com/api-reference/docs/columns#queries
- Docs: https://developer.monday.com/api-reference/docs/docs#queries
- Document Block: https://developer.monday.com/api-reference/docs/blocks#queries

- Files: https://developer.monday.com/api-reference/docs/files#queries
- Folders: https://developer.monday.com/api-reference/docs/folders#queries
- Groups: https://developer.monday.com/api-reference/docs/groups#queries
- Item: https://developer.monday.com/api-reference/docs/items#queries
- Subitems: https://developer.monday.com/api-reference/docs/subitems#queries
- Tags: https://developer.monday.com/api-reference/docs/tags-queries#queries
- Teams: https://developer.monday.com/api-reference/docs/teams#queries
- Updates: https://developer.monday.com/api-reference/docs/updates#queries
- Users: https://developer.monday.com/api-reference/docs/users#queries
- Workspaces: https://developer.monday.com/api-reference/docs/workspaces#queries

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Account	No	No	No	Yes	No
Boards	Yes	Yes	No	Yes	No
Columns	No	No	No	Yes	No
Docs	Yes	Yes	No	Yes	No
Document Blocks	No	Yes	No	Yes	No
Files	Yes	No	No	Yes	No
Groups	No	No	No	Yes	No
Item	Yes	Yes	No	Yes	No
Subitems	No	No	No	Yes	No
Tags	Yes	No	No	Yes	Yes
Teams	Yes	No	No	Yes	No

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Updates	No	Yes	No	Yes	No
Users	Yes	Yes	No	Yes	No
Workspaces	Yes	Yes	No	Yes	No
Folders	Yes	Yes	No	Yes	No

Example

```
monday_read = glueContext.create_dynamic_frame.from_options(
     connection_type="monday",
     connection_options={
         "connectionName": "connectionName",
         "ENTITY_NAME": "account",
         "API_VERSION": "v2"
     }
```

Monday connection option reference

The following are connection options for Monday:

- `ENTITY_NAME`(String) - (Required) Used for Read/Write. The name of your Object in Monday.
- `API_VERSION`(String) - (Required) Used for Read/Write. Monday Rest API version you want to use. Example: v2.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.

Limitations

The following are limitations for the Monday connector:

- The dynamic metadata response has certain conflicts with the documentation as mentioned below:
  - Group, Column entity supports filter operations, but it is not present in the dynamic metadata endpoint, hence it's kept as non-filterable entity.
  - The dynamic endpoint consists of around 15000+ lines and returns metadata of all the entities in a single response, because of this the fields are taking an average of 10 seconds to load hence, this would require some additional time while running a job.
  - Refer the below table for Monday rate limit. The significant size of the dynamic entity's response data causes a noticeable delay, with fields requiring an average of 10 seconds to load.

Complexity Limit	5,000,000 (5M) complexity points
Daily Call Limit	10,000 for Pro Plan
Minute Limit	500 Queries per minutes
Concurrency limit	100 Maximum concurrent requests for Pro Plan

Create a new Monday account:

1. Navigate to Monday’s homepage, https://monday.com/ and choose **Login**.
2. You will be re-directed to the login page. On the bottom of the page, choose **Sign up**.
3. Enter your email address and choose **Continue**. Alternately, you can sign in with Google.
4. Enter the required details and choose **Continue**.
5. Complete the survey questions and follow the steps to complete the account creation process.

Register an OAuth application:

1. Log into your monday.com account. Click on your avatar (picture icon) in the bottom left corner of your screen.
2. Choose **Developer**.
3. Choose **Create app**.
4. Complete the required fields for name and description.

5. Navigate to “OAuth” section present on the right side add the scopes and choose “Save Feature”.
6. Navigate to “Redirect URLS” tab beside the scope and add the redirect URL and choose “Save Feature”.
7. Under the **Redirect URLs** tab, provide the URL of your app. This should be https://{region-code}.console.aws.amazon.com/appflow/oauth. For example, if you are using `us-east-1` you can add `https://us-east-1.console.aws.amazon.com/appflow/oauth`.
8. The application is now ready to use. You can find your credentials, in the “Basic Information” section. Note your Client ID and Client secret strings. These strings are used to make a connection with this app using an AppFlow connector.

Generate personal access token:

Currently, monday.com only offers our V2 API tokens, which are all personal tokens. To access your API tokens, you can use one of two methods depending on your user level. Admin users can utilize both methods to acquire their API tokens. Member users can access their API tokens from their Developer tabs.

Admins - If you are an admin user on your monday.com account,you can access your API tokens from the "Admin" tab with the following steps:

1. Log into your monday.com account. Click on your avatar (picture icon) in the bottom left corner of your screen.
2. Select “Administration” from the resulting menu (this requires you to have admin permissions).
3. Navigate to the “API”Section and generate a “API V2 Token”. You can copy your token and use it.

Developer - If you are a member user on your monday.com account, you can access your API tokens from the Developer tab with the following steps:

1. Log into your monday.com account. Click on your avatar (picture icon) in the bottom left corner of your screen.
2. Select “Developers” from the resulting menu.
3. In the top menu, choose the "Developer" drop-down menu. Select the first option on the drop-down menu titled "My Access Tokens."

Connecting to MongoDB in AWS Glue Studio

AWS Glue provides built-in support for MongoDB. AWS Glue Studio provides a visual interface to connect to MongoDB, author data integration jobs, and run them on the AWS Glue Studio serverless Spark runtime.

Topics

- Creating a MongoDB connection
- Creating a MongoDB source node
- Creating a MongoDB target node
- Advanced options

Creating a MongoDB connection

Prerequisites:

- If your MongoDB instance is in an Amazon VPC, configure Amazon VPC to allow your AWS Glue job to communicate with the MongoDB instance without traffic traversing the public internet.

  In Amazon VPC, identify or create a **VPC**, **Subnet** and **Security group** that AWS Glue will use while executing the job. Additionally, you need to ensure Amazon VPC is configured to permit network traffic between your MongoDB instance and this location. Based on your network layout, this may require changes to security group rules, Network ACLs, NAT Gateways and Peering connections.

To configure a connection to MongoDB:

1. Optionally, in AWS Secrets Manager, create a secret using your MongoDB credentials. To create a secret in Secrets Manager, follow the tutorial available in Create an AWS Secrets Manager secret in the AWS Secrets Manager documentation. After creating the secret, keep the Secret name, *secretName* for the next step.

   - When selecting **Key/value pairs**, create a pair for the key `username` with the value *mongodbUser*.

     When selecting **Key/value pairs**, create a pair for the key `password` with the value *mongodbPass*.

2. In the AWS Glue console, create a connection by following the steps in the section called “Adding an AWS Glue connection”. After creating the connection, keep the connection name, *`connectionName`*, for future use in AWS Glue.

    - When selecting a **Connection type**, select **MongoDB** or **MongoDB Atlas**.
    - When selecting **MongoDB URL** or **MongoDB Atlas URL**, provide the hostname of your MongoDB instance.

      A MongoDB URL is provided in the format `mongodb://`*`mongoHost`*`:`*`mongoPort`*`/`*`mongoDBname`*.

      A MongoDB Atlas URL is provided in the format `mongodb+srv://`*`mongoHost`*`:`*`mongoPort`*`/`*`mongoDBname`*.

      Providing the default database for the connection, *`mongoDBname`* is optional.
    - If you chose to create an Secrets Manager secret, choose the AWS Secrets Manager **Credential type**.

      Then, in **AWS Secret** provide *`secretName`*.
    - If you choose to provide **Username and password**, provide *`mongodbUser`* and *`mongodbPass`*.

3. In the following situations, you may require additional configuration:

    - For MongoDB instances hosted on AWS in an Amazon VPC
        - You will need to provide Amazon VPC connection information to the AWS Glue connection that defines your MongoDB security credentials. When creating or updating your connection, set **VPC**, **Subnet** and **Security groups** in **Network options**.

After creating a AWS Glue MongoDB connection, you will need to perform the following steps before running your AWS Glue job:

- When working with AWS Glue jobs in the visual editor, you must provide Amazon VPC connection information for your job to connect to MongoDB. Identify a suitable location in Amazon VPC and provide it to your AWS Glue MongoDB connection.
- If you chose to create an Secrets Manager secret, grant the IAM role associated with your AWS Glue job permission to read *`secretName`*.

Creating a MongoDB source node

Prerequisites needed

- A AWS Glue MongoDB connection, as described in the previous section, the section called "Creating a MongoDB connection".
- If you chose to create an Secrets Manager secret, appropriate permissions on your job to read the secret used by the connection.
- A MongoDB collection you would like to read from. You will need identification information for the collection.

  A MongoDB collection is identified by a database name and a collection name, *`mongodbName`*, *`mongodbCollection`*.

Adding a MongoDB data source

To add a Data source – MongoDB node:

1. Choose the connection for your MongoDB data source. Since you have created it, it should be available in the dropdown. If you need to create a connection, choose **Create MongoDB connection**. For more information see the previous section, the section called "Creating a MongoDB connection".

   Once you have chosen a connection, you can view the connection properties by clicking **View properties**.
2. Choose a **Database**. Enter *`mongodbName`*.
3. Choose a **Collection**. Enter *`mongodbCollection`*.
4. Choose your **Partitioner**, **Partition size (MB)** and **Partition key**. For more information about partition parameters, see the section called ""connectionType": "mongodb" as source".
5. In **Custom MongoDB properties**, enter parameters and values as needed.

Creating a MongoDB target node

Prerequisites needed

- A AWS Glue MongoDB connection, configured with an AWS Secrets Manager secret, as described in the previous section, the section called "Creating a MongoDB connection".

- Appropriate permissions on your job to read the secret used by the connection.
- A MongoDB table you would like to write to, *`tableName`*.

Adding a MongoDB data target

To add a Data target – MongoDB node:

1. Choose the connection for your MongoDB data source. Since you have created it, it should be available in the dropdown. If you need to create a connection, choose **Create MongoDB connection**. For more information see the previous section, the section called "Creating a MongoDB connection".

   Once you have chosen a connection, you can view the connection properties by clicking **View properties**.
2. Choose a **Database**. Enter *`mongodbName`*.
3. Choose a **Collection**. Enter *`mongodbCollection`*.
4. Choose your **Partitioner**, **Partition size (MB)** and **Partition key**. For more information about partition parameters, see the section called ""connectionType": "mongodb" as source".
5. Choose **Retry Writes** if desired.
6. In **Custom MongoDB properties**, enter parameters and values as needed.

Advanced options

You can provide advanced options when creating a MongoDB node. These options are the same as those available when programming AWS Glue for Spark scripts.

See the section called "MongoDB connection".

Connecting to Oracle NetSuite

Oracle NetSuite is an all-in-one cloud business management solution that helps organizations operate more effectively by automating core processes and providing real-time visibility into operational and financial performance. With a single, integrated suite of applications for managing accounting, order processing, inventory management, production, supply chain and warehouse operations, Oracle NetSuite gives companies clear visibility into their data and tighter control over their businesses.

Topics

- AWS Glue support for Oracle NetSuite
- Policies containing the API operations for creating and using connections
- Configuring Oracle NetSuite
- Configuring Oracle NetSuite connections
- Reading from Oracle NetSuite entities
- Oracle NetSuite connection options
- Limitations and notes for Oracle NetSuite connector

AWS Glue support for Oracle NetSuite

AWS Glue supports Oracle NetSuite as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Oracle NetSuite.

Supported as a target?

No.

Supported Oracle NetSuite API versions

The following Oracle NetSuite API versions are supported:

- v1

For entity support per version specific, see Supported entities for Source.

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Oracle NetSuite

Before you can use AWS Glue to transfer data from Oracle NetSuite, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- You have a Oracle NetSuite account. For more information, see Creating a Oracle NetSuite account.
- Your Oracle NetSuite account is enabled for API access.

- You have created an OAuth 2.0 API integration in your Oracle NetSuite developer account. This integration provides the client credentials that AWS Glue uses to access your data securely when it makes authenticated calls to your account. For more information, see Creating a Oracle NetSuite client app and OAuth 2.0 credentials.

If you meet these requirements, you're ready to connect AWS Glue to your Oracle NetSuite account.

Creating a Oracle NetSuite account

Navigate to Oracle NetSuite, and choose **Free Product Tour**. Fill in the required details to get a free product tour, through which you can contact a vendor. The process for procuring an account is as follows:

- The procurement of a NetSuite account is done via a vendor, who provides a form/quote which has to be legally reviewed.
- The account to be procured for Oracle NetSuite connector is of **Standard Cloud Service**.
- This account is created by the vendor and temporary credentials are shared by them. You will receive a welcome mail from NetSuite <billing@notification.netsuite.com> <system@sent-via.netsuite.com> with the details such as your username, and a link to set your password.
- Use the **Set your password** link to set the password for the username provided by the vendor.

Creating a Oracle NetSuite client app and OAuth 2.0 credentials

To get the Client ID and Client Secret you create a Oracle NetSuite client app:

1. Log into your NetSuite account through the NetSuite customer login.
2. Choose **Setup** > **Company** > **Enable features**.
3. Navigate to the **SuiteCloud** section and select the **REST WEB SERVICES** checkbox under **SuiteTalk (Web Services)**.
4. Select the **OAUTH 2.0** checkbox under **Manage Authentication**. Click **Save**.
5. Go to **Setup** > **Integration** > **Manage Integrations** and choose **New** to create an OAuth2.0 application.
6. Enter a name of your choice and keep the **STATE** as Enabled.
7. If checked, uncheck the **TBA: AUTHORIZATION FLOW** and **TOKEN-BASED AUTHENTICATION** checkboxes displayed under **Token-based Authentication**.

8. Select the **AUTHORIZATION CODE GRANT** and **PUBLIC CLIENT** checkboxes under **OAuth 2.0**.
9. Under Auth, note the Client ID and Client Secret.
10. Enter a **REDIRECT URI**. For example, https://us-east-1.console.aws.amazon.com/gluestudio/oauth
11. Select the **REST WEB SERVICES** checkbox under **SCOPE**.
12. Select the **USER CREDENTIALS** checkbox under **User Credentials**. Choose **Save**.
13. Note the CONSUMER KEY/CLIENT ID and CONSUMER SECRET/CLIENT SECRET under **Client Credentials**. These values are displayed only once.
14. Create an ADMINISTRATOR role if needed by navigating to **User/Roles** > **Manage Roles** > **New**.
15. While creating a custom role, add full access under the **Permissions** tab for the following entities/functionalities:
    - "Deposit", "Items", "Item Fulfillment", "Make Journal Entry", "Purchase Order", "Subsidiaries", "Vendors", "Bills", "Vendor Return Authorization", "Track Time", "Customer Payment", "Custom Record Entries", "Custom Record Types", "REST Web Services", "OAuth 2.0 Authorized Applications Management", "Custom Entity Fields", "Log in using OAuth 2.0 Access Tokens".

For more information see OAuth 2.0 in the NetSuite Applications Suite documentation.

Configuring Oracle NetSuite connections

Oracle NetSuite supports the AUTHORIZATION_CODE grant type for OAuth2. The grant type determines how AWS Glue communicates with Oracle NetSuite to request access to your data.

- This grant type is considered "three-legged" OAuth as it relies on redirecting users to a third-party authorization server to authenticate the user. It is used when creating connections via the AWS Glue console. The user creating a connection may by default rely on an AWS Glue-owned connected app (AWS Glue managed client application) where they do not need to provide any OAuth-related information except for their Oracle NetSuite instance URL. The AWS Glue console will redirect the user to Oracle NetSuite where the user must log in and allow AWS Glue the requested permissions to access their Oracle NetSuite instance.
- Users may still opt to create their own connected app in Oracle NetSuite and provide their own client id and client secret when creating connections through the AWS Glue console. In this scenario, they will still be redirected to Oracle NetSuite to login and authorize AWS Glue to access their resources.

- This grant type results in a refresh token and access token. The access token is short lived, and may be refreshed automatically without user interaction using the refresh token.
- For public Oracle NetSuite documentation on creating a connected app for Authorization Code OAuth flow, see Public apps.

To configure a Oracle NetSuite connection:

1. In AWS Secrets Manager, create a secret with the following details:
   a. For the customer managed connected app, the Secret should contain the connected app Consumer Secret with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as key.
   b. Note: You must create a secret for your connection in AWS Glue.
2. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Connection type**, select Oracle NetSuite.
   b. Provide the Oracle NetSuite environment.
   c. Select the AWS IAM role which AWS Glue can assume and has permissions for following actions:

      JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

    d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.
    e. Select the network options if you want to use your network.

3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Oracle NetSuite entities

Prerequisite

A Oracle NetSuite object you would like to read from. You will need the object name such as `deposit` or `timebill`. The following table shows the supported entities.

Supported entities for source:

Entity	Can be filtered	Supports Order By	Supports Limit	Supports SELECT *	Supports partitioning
Deposit	Yes	No	Yes	Yes	Yes
Description Item	Yes	No	Yes	Yes	Yes
Inventory Item	Yes	No	Yes	Yes	Yes
Item Fulfillment	Yes	No	Yes	Yes	Yes
Item Group	Yes	No	Yes	Yes	Yes
Journal Entry	Yes	No	Yes	Yes	Yes
Non-Inventory Purchase Item	Yes	No	Yes	Yes	Yes

Entity	Can be filtered	Supports Order By	Supports Limit	Supports SELECT *	Supports partitioning
Non-Inventory Resale Item	Yes	No	Yes	Yes	Yes
Non-Inventory Sale Item	Yes	No	Yes	Yes	Yes
Purchase Order	Yes	No	Yes	Yes	Yes
Subsidiary	Yes	No	Yes	Yes	Yes
Vendor	Yes	No	Yes	Yes	Yes
Vendor Bill	Yes	No	Yes	Yes	Yes
Vendor Return Authorization	Yes	No	Yes	Yes	Yes
Time Bill	Yes	No	Yes	Yes	Yes
Customer Payment	Yes	No	Yes	Yes	Yes
Fulfillment Request	Yes	No	Yes	Yes	Yes
Item	Yes	Yes	Yes	Yes	Yes
Transaction Line	Yes	Yes	Yes	Yes	Yes
Transaction Accounting Line	Yes	Yes	Yes	Yes	Yes

Entity	Can be filtered	Supports Order By	Supports Limit	Supports SELECT *	Supports partitioning
Custom Record Types (Dynamic)	Yes	Yes	Yes	Yes	Yes

Example:

```
netsuiteerp_read = glueContext.create_dynamic_frame.from_options(
    connection_type="netsuiteerp",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "deposit",
        "API_VERSION": "v1"
    }
)
```

Oracle NetSuite entity and field details:

Oracle NetSuite dynamically loads available fields under selected entity. Depending on the data type of the field, it supports the following filter operators.

Field data type	Supported filter operators
String	LIKE, =, !=
Date	BETWEEN, =, <, <=, >, >=
DateTime	BETWEEN, <, <=, >, >=
Numeric	=, !=, <, <=, >, >=
Boolean	=, !=

Expected input format for Boolean values in Filter Expression:

Entity	Boolean "true" value format	Boolean "false" value format	Example
Item, Transaction Line, Transaction Accounting Line, and Custom Record Type entities	T or t	F or f	isinactive = "T" or isinactive = "t"
All other entities	true	false	isinactive = true

Partitioning queries

Field-based partitioning

The Oracle NetSuite connector has dynamic metadata so that supported fields for field based partitioning are chosen dynamically. Field based partitioning is supported on fields having the data type Integer, BigInteger, Date or DateTime.

You can provide the additional Spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, and `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition the query.
- `LOWER_BOUND`: an **inclusive** lower bound value of the chosen partition field.

  For the timestamp field, we accept the Spark timestamp format used in Spark SQL queries.

  Examples of valid values:

  ```
  "TIMESTAMP \"1707256978123\""
  "TIMESTAMP \"1702600882\""
  "TIMESTAMP '2024-02-06T22:00:00:00.000Z'"
  "TIMESTAMP '2024-02-06T22:00:00:00Z'"
  "TIMESTAMP '2024-02-06'"
  ```
- `UPPER_BOUND`: an **exclusive** upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: the number of partitions.

Example:

```
netsuiteerp_read = glueContext.create_dynamic_frame.from_options(
    connection_type="netsuiteerp",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "deposit",
        "API_VERSION": "v1",
        "PARTITION_FIELD": "id",
        "LOWER_BOUND": "1",
        "UPPER_BOUND": "10000",
        "NUM_PARTITIONS": "10"
    }
```

Record-based partitioning

You can provide the additional Spark option `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With this parameter, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

In record based partitioning, the total number of records present is queried from Oracle NetSuite API, and it is divided by the `NUM_PARTITIONS` number provided, the resulting number of records are then concurrently fetched by each sub-query.

- `NUM_PARTITIONS`: the number of partitions.

Example:

```
netsuiteerp_read = glueContext.create_dynamic_frame.from_options(
    connection_type="netsuiteerp",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "deposit",
        "API_VERSION": "v1",
        "NUM_PARTITIONS": "3"
    }
```

Oracle NetSuite connection options

The following are connection options for Oracle NetSuite:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of the Oracle NetSuite entity. Example: deposit.
- `API_VERSION`(String) - (Required) Used for Read. Oracle NetSuite Rest API version you want to use. The value will be v1, as Oracle NetSuite currently supports only version v1.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Comma-separated list of columns you want to select for the selected entity.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition the query (field-based partitioning).
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field, used in field-based partitioning.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field, used in field-based partitioning.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read. Used in both field- and record- based partitioning.
- `INSTANCEE_URL`(String) - A valid NetSuite instance URL with format https://{account-id}.suitetalk.api.netsuite.com.

Limitations and notes for Oracle NetSuite connector

The following are limitations or notes for the Oracle NetSuite connector:

- The values of the access_token and refresh_token parameters are in JSON Web Token (JWT) format. The access token is valid for 60 minutes whereas the refresh_token is valid for seven days.
- During client ID and client secret generation, if you select "PUBLIC CLIENT" along with "AUTHORIZATION CODE GRANT", then the refresh token is only valid for three hours and is for one-time use only.
- You can fetch at most 1,00,000 records using the connector. For more information, refer to Executing SuiteQL Queries Through REST Web Services.
- Partitions are created such that each partition will fetch records in multiples of 1000, except possibly the last one which will fetch the remaining records.

- For Item, Transaction Line and Transaction Accounting Line objects, the connector will not support a few operators due to following reasons:
  - Applying the `EQUAL_TO`, `NOT_EQUAL_TO` filter operators to fields of type Date gives unreliable results.
  - Applying the `LESS_THAN_OR_EQUAL_TO` filter operator to fields of type Date gives unreliable results and behaves similar to the `LESS_THAN` operator.
  - Applying the `GREATER_THAN` filter operator to fields of type Date= gives unreliable results and behaves similar to `GREATER_THAN_OR_EQUAL_TO` operator.
- For Item, Transaction Line, Transaction Accounting Line, and Custom Record Type objects, boolean values come in the format T/F instead of the standard true/false. The connector maps the t/f values to true/false to ensure consistency in data.

Connecting to OpenSearch Service in AWS Glue Studio

AWS Glue provides built-in support for Amazon OpenSearch Service. AWS Glue Studio provides a visual interface to connect to Amazon OpenSearch Service, author data integration jobs, and run them on the AWS Glue Studio serverless Spark runtime. This feature is not compatible with OpenSearch Service serverless.

AWS Glue Studio creates a unified connection for Amazon OpenSearch Service. For more information, see Considerations.

Topics

- Creating a OpenSearch Service connection
- Creating a OpenSearch Service source node
- Creating a OpenSearch Service target node
- Advanced options

Creating a OpenSearch Service connection

Prerequisites:

- Identify the domain endpoint, *`aosEndpoint`* and port, *`aosPort`* you would like to read from, or create the resource by following instructions in the Amazon OpenSearch Service documentation. For more information on creating a domain, see Creating and managing Amazon OpenSearch Service domains in the Amazon OpenSearch Service documentation.

An Amazon OpenSearch Service domain endpoint will have the following default form, https://search-*domainName*-*unstructuredIdContent*.*region*.es.amazonaws.com. For more information on identifying your domain endpoint, see Creating and managing Amazon OpenSearch Service domains in the Amazon OpenSearch Service documentation.

Identify or generate HTTP basic authentication credentials, *aosUser* and *aosPassword* for your domain.

To configure a connection to OpenSearch Service:

1. In AWS Secrets Manager, create a secret using your OpenSearch Service credentials. To create a secret in Secrets Manager, follow the tutorial available in Create an AWS Secrets Manager secret in the AWS Secrets Manager documentation. After creating the secret, keep the Secret name, *secretName* for the next step.
    - When selecting **Key/value pairs**, create a pair for the key `USERNAME` with the value *aosUser*.
    - When selecting **Key/value pairs**, create a pair for the key `PASSWORD` with the value *aosPassword*.
2. In the AWS Glue console, create a connection by following the steps in the section called "Adding an AWS Glue connection". After creating the connection, keep the connection name, *connectionName*, for future use in AWS Glue.
    - When selecting a **Connection type**, select OpenSearch Service.
    - When selecting a Domain endpoint, provide *aosEndpoint*.
    - When selecting a port, provide *aosPort*.
    - When selecting an **AWS Secret**, provide *secretName*.

Creating a OpenSearch Service source node

Prerequisites needed

- A AWS Glue OpenSearch Service connection, configured with an AWS Secrets Manager secret, as described in the previous section, the section called "Creating a OpenSearch Service connection".
- Appropriate permissions on your job to read the secret used by the connection.
- A OpenSearch Service index you would like to read from, *aosIndex*.

Adding a OpenSearch Service data source

To add a Data source – OpenSearch Service node:

1. Choose the connection for your OpenSearch Service data source. Since you have created it, it should be available in the dropdown. If you need to create a connection, choose **Create OpenSearch Service connection**. For more information see the previous section, the section called "Creating a OpenSearch Service connection".

   Once you have chosen a connection, you can view the connection properties by clicking **View properties**.
2. Provide **Index**, the index you would like to read.
3. Optionally, provide **Query**, an OpenSearch query to deliver more specific results. For more information about writing OpenSearch queries, consult the section called "Read from OpenSearch Service".
4. In **Custom OpenSearch Service properties**, enter parameters and values as needed.

Creating a OpenSearch Service target node

Prerequisites needed

- A AWS Glue OpenSearch Service connection, configured with an AWS Secrets Manager secret, as described in the previous section, the section called "Creating a OpenSearch Service connection".
- Appropriate permissions on your job to read the secret used by the connection.
- A OpenSearch Service index you would like to write to, *`aosIndex`*.

Adding a OpenSearch Service data target

To add a Data target – OpenSearch Service node:

1. Choose the connection for your OpenSearch Service data source. Since you have created it, it should be available in the dropdown. If you need to create a connection, choose **Create OpenSearch Service connection**. For more information see the previous section, the section called "Creating a OpenSearch Service connection".

   Once you have chosen a connection, you can view the connection properties by clicking **View properties**.

2. Provide **Index**, the index you would like to read.
3. In **Custom OpenSearch Service properties**, enter parameters and values as needed.

Advanced options

You can provide advanced options when creating a OpenSearch Service node. These options are the same as those available when programming AWS Glue for Spark scripts.

See the section called "OpenSearch Service connections".

Connecting to Okta

The Okta API is the programmatic interface to Okta, used for managing large or complex Okta accounts and campaigns. If you're a Okta user, you can connect AWS Glue to your Okta account. Then, you can use Okta as a data source in your ETL jobs. Run these jobs to transfer data between Okta and AWS services or other supported applications.

Topics

- AWS Glue support for Okta
- Policies containing the API operations for creating and using connections
- Configuring Okta
- Configuring Okta connections
- Reading from Okta entities
- Okta connection option reference
- Okta New Account and Developer App creation steps
- Limitations

AWS Glue support for Okta

AWS Glue supports Okta as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Okta.

Supported as a target?

No.

Supported Okta API versions

v1.

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```json
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

You can also use the following managed IAM policies to allow access:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Okta

Before you can use AWS Glue to transfer data to or from Okta, you must meet these requirements:

Minimum requirements

- You have a Okta account. For more information on creating an account, see Okta New Account and Developer App creation steps.
- Your Okta account is enabled for API access.
- You have created a OAuth2 API integration in your Okta account. This integration provides the client credentials that AWS Glue uses to access your data securely when it makes authenticated calls to your account. For more information, refer Steps to Create a Client app and OAuth2.0 credentials: Okta New Account and Developer App Creation Steps
- You have a Okta account with a OktaApiToken. Refer to Okta documentation .

If you meet these requirements, you're ready to connect AWS Glue to your Okta account. For typical connections, you don't need do anything else in Okta.

Configuring Okta connections

Okta supports two types of authentication mechanisms:

- OAuth auth: Okta supports the `AUTHORIZATION_CODE` grant type.
  - This grant type is considered “three-legged” OAuth as it relies on redirecting users to the third party authorization server to authenticate the user. It is used when creating connections via the AWS Glue Console. The AWS Glue Console will redirect the user to Okta where the user must login and allow AWS Glue the requested permissions to access their Okta instance.
  - Users may opt to create their own connected app in Okta and provide their own client ID and client secret when creating connections through the AWS Glue Console. In this scenario, they will still be redirected to Okta to login and authorize AWS Glue to access their resources.
  - This grant type results in a refresh token and access token. The access token is short lived, and may be refreshed automatically without user interaction using the refresh token.
  - For more information, see public Okta documentation on creating a connected app for Authorization Code OAuth flow .
- Custom auth:
  - For public Okta documentation on generating the required API keys for custom authorization, see Okta documentation .

To configure an Okta connection:

1. In AWS Secrets Manager, create a secret with the following details. It is required to create a secret for each connection in AWS Glue.
    a. For OAuth auth:
        - For customer managed connected app – Secret should contain the connected app Consumer Secret with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as key.
    b. For Custom auth:
        - For customer managed connected app - Secret should contain the connected app Consumer Secret with `OktaApiToken` as key.
2. In AWS Glue Studio, create a connection under **Data Connections** by following the steps below:
    a. Under Connections, choose **Create connection**.
    b. When selecting a **Data Source**, select Okta.
    c. Provide your Okta subdomain.
    d. Select the Okta Domain URL of your Okta account.
    e. Select the IAM role which AWS Glue can assume and has permissions for following actions:

    JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

    f. Select Authentication Type to connect to data source.

g. For OAuth2 authentication type, provide the **User Managed Client Application ClientId** of the Okta app.

h. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.

i. Select the network options if you want to use your network.

3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.
4. In your AWS Glue job configuration, provide `connectionName` as an **Additional network connection**.

Reading from Okta entities

Prerequisites

- A Okta Object you would like to read from. Refer the supported entities table below to check the available entities.

Supported entities

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Applications	Yes	Yes	No	Yes	No
Devices	Yes	Yes	No	Yes	Yes
Groups	Yes	Yes	Yes	Yes	Yes
Users	Yes	Yes	Yes	Yes	Yes
User Types	No	No	No	Yes	No

Example

```
okta_read = glueContext.create_dynamic_frame.from_options(
    connection_type="Okta",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "applications",
        "API_VERSION": "v1"
    }
```

Okta entity and field details

Entities list:

- Application: https://developer.okta.com/docs/api/openapi/okta-management/management/tag/Application/
- Device: https://developer.okta.com/docs/api/openapi/okta-management/management/tag/Device/
- Group: https://developer.okta.com/docs/api/openapi/okta-management/management/tag/Group/
- User: https://developer.okta.com/docs/api/openapi/okta-management/management/tag/User/
- User Type: https://developer.okta.com/docs/api/openapi/okta-management/management/tag/UserType/

Partitioning queries

Additional spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, `NUM_PARTITIONS` can be provided if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition query.
- `LOWER_BOUND`: an inclusive lower bound value of the chosen partition field.

  For date, we accept the Spark date format used in Spark SQL queries. Example of valid values: `"2024-02-06"`.
- `UPPER_BOUND`: an exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: number of partitions.

Example

```
okta_read = glueContext.create_dynamic_frame.from_options(
    connection_type="okta",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "lastUpdated",
        "API_VERSION": "v1",
        "PARTITION_FIELD": "lastMembershipUpdated"
        "LOWER_BOUND": "2022-08-10T10:28:46.000Z"
        "UPPER_BOUND": "2024-08-10T10:28:46.000Z"
        "NUM_PARTITIONS": "10"
    }
```

Okta connection option reference

The following are connection options for Okta:

- `ENTITY_NAME`(String) - (Required) Used for Read/Write. The name of your Object in Okta.
- `API_VERSION`(String) - (Required) Used for Read/Write. Okta Rest API version you want to use. Example: v1.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.

Okta New Account and Developer App creation steps

Create a developer account on Okta for getting access to the Okta API. A free Okta developer account gives access to most key developer features required for accessing Okt API.

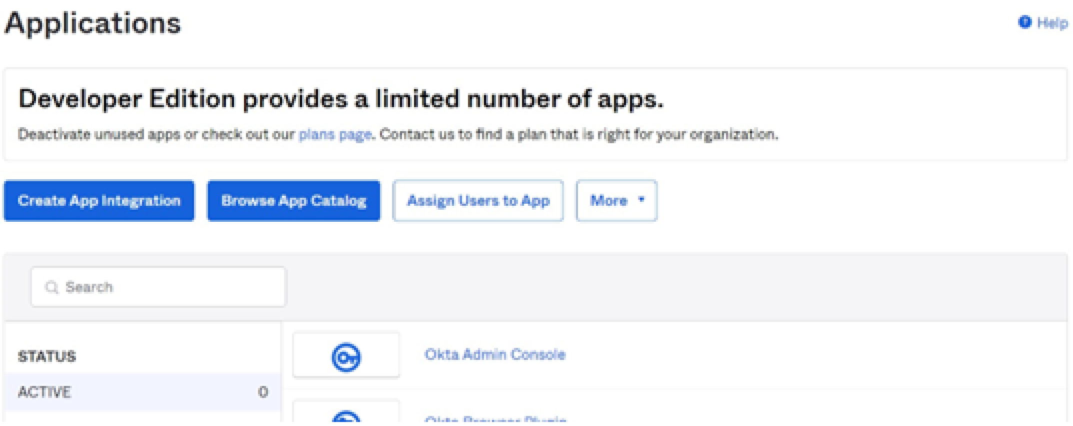
To create a developer account on Okta

1. Navigate to https://developer.okta.com/signup/.

2. Enter account information email, first name, last name and country/region. Choose **I’m not a robot** and then, **Signup**.
3. A verification mail is sent to your registered mail id. You will receive a link in your email for activating your Okta developer account. Choose **Activate**.
4. You will be redirected to password reset page. Enter the new password twice and choose **Reset password**.
5. You will be redirected to your Okta developer account dashboard.

To create a client app and OAuth 2.0 credentials

1. In the developer dashboard, choose create app integration.



Developer Edition provides a limited number of apps.

Deactivate unused apps or check out our plans page. Contact us to find a plan that is right for your organization.

2. The **Create a new app Integration** window will appear and present various sign-in methods. Select **OIDC –OpenID Connect**.
3. Scroll down to the Application type section. Select as a **Web Application** and choose **Next**.
4. On the "New Web App Integration" screen, fill following information:
    - App integration name - Enter the name of the app.
    - Grant type - Choose **Authorization Code** and **Refresh Token** from the list.
    - Sign-in redirect URIs - Choose **Add URI** and add `https://{regioncode}.console.aws.amazon.com/appflow/oauth`. For example, if you are using `us-west-2 (Oregon)` you can add `https://us-east-1.console.aws.amazon.com/appflow/oauth`.
    - Controlled Access - Assign the app to your user groups as you require and choose **Save**.
5. Your Client Id and Client Secret is generated.

Limitations

The following are limitations for the Okta connector:

- For ‘Applications’ entity only one filter can be applied. If more than 1 filter is applied then 400 Bad Request is return with error summary –‘Invalid Search criteria’.
- Order by can be supported with search queries only. For example, `http://dev-15940405.okta.com/api/v1/groups?search=type e.q. "OKTA_GROUP"&sortBy=lastUpdated&sortOrder=asc`

Connecting to PayPal

PayPal is a payments system that facilitates online money transfers between parties, such as transfers between customers and online vendors. If you're a PayPal user, your account contains data about your transactions, such as their payers, dates, and statuses. You can use AWS Glue to transfer data from PayPal to certain AWS services or other supported applications.

Topics

- AWS Glue support for PayPal
- Policies containing the API operations for creating and using connections
- Configuring PayPal
- Configuring PayPal connections
- Reading from PayPal entities
- PayPal connection options
- Limitations and notes for PayPal connector

AWS Glue support for PayPal

AWS Glue supports PayPal as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from PayPal.

Supported as a target?

No.

Supported PayPal API versions

The following PayPal API versions are supported:

- v1

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention

for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring PayPal

Before you can use AWS Glue to transfer data from PayPal, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- You have a PayPal account with client credentials.
- Your PayPal account has API access with a valid license.

If you meet these requirements, you’re ready to connect AWS Glue to your PayPal account. For typical connections, you don't need do anything else in PayPal.

Configuring PayPal connections

PayPal supports the CLIENT CREDENTIALS grant type for OAuth2.

- This grant type is considered 2-legged OAuth 2.0 as it is used by clients to obtain an access token outside of the context of a user. AWS Glue is able to use the client ID and client secret to authenticate the PayPal APIs which are provided by custom services that you define.
- Each custom service is owned by an API-only user which has a set of roles and permissions which authorize the service to perform specific actions. An access token is associated with a single custom service.
- This grant type results in an access token which is short lived, and may be renewed by calling the `/v2/oauth2/token` endpoint again.
- For public PayPal documentation for OAuth 2.0 with client credentials, see Authentication.

To configure a PayPal connection:

1. In AWS Secrets Manager, create a secret with the following details:
   a. For the customer managed connected app, the Secret should contain the connected app Consumer Secret with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as key.
   b. Note: you must create a secret for your connections in AWS Glue.

1. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
    a. When selecting a **Connection type**, select PayPal.
    b. Provide the `INSTANCE_URL` of the PayPal instance you want to connect to.
    c. Select the AWS IAM role which AWS Glue can assume and has permissions for following actions:

       JSON

       ```
       {
         "Version": "2012-10-17",
         "Statement": [
           {
             "Effect": "Allow",
             "Action": [
               "secretsmanager:DescribeSecret",
               "secretsmanager:GetSecretValue",
               "secretsmanager:PutSecretValue",
               "ec2:CreateNetworkInterface",
               "ec2:DescribeNetworkInterfaces",
               "ec2:DeleteNetworkInterface"
             ],
             "Resource": "*"
           }
         ]
       }
       ```

    d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.
    e. Select the network options if you want to use your network.
2. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Getting OAuth 2.0 credentials

To call the Rest API, you'll need to exchange your client ID and client secret for an access token. For more information, see Get started with PayPal REST APIs .

Reading from PayPal entities

Prerequisite

A PayPal object you would like to read from. You will need the object name, `transaction`.

Supported entities for source:

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
transaction	Yes	Yes	No	Yes	Yes

Example:

```
paypal_read = glueContext.create_dynamic_frame.from_options(
    connection_type="paypal",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "transaction",
        "API_VERSION": "v1",
        "INSTANCE_URL": "https://api-m.paypal.com"
    }
```

PayPal entity and field details:

Entities with static metadata:

Entity	Field	Data type	Supported operators
transaction	transaction_initiation_date	DateTime	Between
	last_refreshed_datetime	String	N/A
	payment_instrument_type	String	=
	balance_affecting_records_only	String	=
	store_id	String	=

Entity	Field	Data type	Supported operators
	terminal_id	String	=
	transaction_currency	String	=
	transaction_id	String	N/A
	transaction_status	String	N/A
	transaction_type	String	N/A
	transaction_info	Struct	N/A
	payer_info	Struct	N/A
	shipping_info	Struct	N/A
	cart_info	Struct	N/A
	store_info	Struct	N/A
	auction_info	Struct	N/A
	incentive_info	Struct	N/A

Partitioning queries

You can provide the additional Spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, and `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition the query.
- `LOWER_BOUND`: an **inclusive** lower bound value of the chosen partition field.

  For the Datetime field, we accept the value in ISO format.

  Examples of valid value:

```
"2024-07-01T00:00:00.000Z"
```

- `UPPER_BOUND`: an **exclusive** upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: the number of partitions.

The following field is supported for entity-wise partitioning:

Entity name	Partitioning fields	Data type
transaction	transaction_initiation_date	DateTime

Example:

```
paypal_read = glueContext.create_dynamic_frame.from_options(
    connection_type="paypal",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "transaction",
        "API_VERSION": "v1",
        "PARTITION_FIELD": "transaction_initiation_date"
        "LOWER_BOUND": "2024-07-01T00:00:00.000Z"
        "UPPER_BOUND": "2024-07-02T00:00:00.000Z"
        "NUM_PARTITIONS": "10"
    }
```

PayPal connection options

The following are connection options for PayPal:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your object in PayPal.
- `API_VERSION`(String) - (Required) Used for Read. PayPal Rest API version you want to use.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.

- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.

Limitations and notes for PayPal connector

The following are limitations or notes for the PayPal connector:

- The PayPal transactions documentation mentions that it takes a maximum of three hours for executed transactions to appear in the list transactions call. However, it has been observed to take more time than that depending the `last_refreshed_datetime`. Here, `last_refreshed_datetime` is the amount of time until which you have data available from the APIs.
- If the `last_refreshed_datetime` is less than the requested `end_date` then, the `end_date` becomes equal to the `last_refreshed_datetime` as we only have data up until that point.
- The `transaction_initiation_date` field is a mandatory filter to be provided for the `transaction` entity and the maximum supported date range for this field is 31 days.
- When you call the `transaction` entity API request with filters (query parameters) other than the `transaction_initiation_date` field, it is expected that the value of the `ending_balance` field won't be fetched in the response.

Connecting to Pendo

Pendo provides a rich data store for user interaction data. Customers will transfer this data to AWS so that they may join it with other product data, perform additional analysis and dash-boarding and set alerts if they choose.

Topics

- AWS Glue support for Pendo
- Policies containing the API operations for creating and using connections
- Configuring Pendo
- Configuring Pendo connections

- Reading from Pendo entities
- Pendo connection options
- Limitations

AWS Glue support for Pendo

AWS Glue supports Pendo as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Pendo.

Supported as a target?

No.

Supported Pendo API versions

v1

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
    ]
}
```

If you don't want to use the preceding method, alternatively, use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Pendo

Before you can use AWS Glue to transfer from Pendo, you must meet the following requirements:

Minimum requirements

- You have a Pendo account with an `apiKey` with `write access` enabled.
- Your Pendo account has API access with a valid license.

If you meet these requirements, you’re ready to connect AWS Glue to your Pendo account. For typical connections, you don't need do anything else in Pendo.

Configuring Pendo connections

Pendo supports custom authentication.

For public Pendo documentation on generating the required API keys for custom authorization, refer Authentication – Pendo REST API Documentation

To configure a Pendo connection:

1. In AWS Secrets Manager, create a secret with the following details:

- For customer managed connected app - Secret should contain the connected app Consumer Secret with `apiKey` as the key.

> **Note**
>
> It is a must to create a secret per connection in AWS Glue.

2. In AWS Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Data Source**, select Pendo.
   b. Provide the `instanceUrl` of the Pendo instance you want to connect to.
   c. Select the IAM role for which AWS Glue can assume and has permissions for following actions:

      JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

   d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.
   e. Select the network options if you want to use your network.
3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.
4. In your AWS Glue job configuration, provide `connectionName` as an Additional network connection.

Reading from Pendo entities

Prerequisites

An Pendo Object you would like to read from. Refer the supported entities table below to check the available entities.

Supported entities

- Feature
- Guide
- Page
- Report
- Report Data
- Visitor
- Account
- Event
- Feature Event
- Guide Event
- Page Event
- Poll Event
- Track Event

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Feature	No	No	No	Yes	No
Guide	No	No	No	Yes	No
Page	No	No	No	Yes	No
Report	No	No	No	Yes	No
Report Data	No	No	No	Yes	No

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Visitor (Aggregation API)	Yes	No	Yes	Yes	No
Account (Aggregation API)	Yes	No	Yes	Yes	No
Event (Aggregation API)	Yes	No	Yes	Yes	No
Feature Event (Aggregation API)	Yes	No	Yes	Yes	Yes
Guide Event (Aggregation API)	Yes	No	Yes	Yes	Yes
Account (Aggregation API)	Yes	No	Yes	Yes	Yes
Page Event (Aggregation API)	Yes	No	Yes	Yes	Yes
Poll Event (Aggregation API)	Yes	No	Yes	Yes	Yes
Track Event (Aggregation API)	Yes	No	Yes	Yes	Yes

Example

```
Pendo_read = glueContext.create_dynamic_frame.from_options(
    connection_type="glue.spark.Pendo",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "feature",
        "API_VERSION": "v1",
        "INSTANCE_URL": "instanceUrl"
    }
```

Partitioning queries

You can provide the additional Spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, and `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition the query.
- `LOWER_BOUND`: an **inclusive** lower bound value of the chosen partition field.

  For the DateTime field, we accept the value in ISO format.

  Example of valid value:

  ```
  "2024-07-01T00:00:00.000Z"
  ```
- `UPPER_BOUND`: an **exclusive** upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: the number of partitions.

The following table describes the entity partitioning field support details:

Entity name
Event
Feature Event
Guide Event

Entity name
Page Event
Poll Event
Track Event

Example:

```
pendo_read = glueContext.create_dynamic_frame.from_options(
    connection_type="glue.spark.pendo",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "event",
        "API_VERSION": "v1",
        "INSTANCE_URL": "instanceUrl"
        "NUM_PARTITIONS": "10",
        "PARTITION_FIELD": "appId"
        "LOWER_BOUND": "4656"
        "UPPER_BOUND": "7788"
    }
```

Pendo connection options

The following are connection options for Pendo:

- `ENTITY_NAME`(String) – (Required) Used for Read/Write. The name of your Object in Pendo.
- `INSTANCE_URL`(String) - (Required) A valid Pendo Instance URL with the following allowed values:
  - Default
  - Europe
  - US1
- `API_VERSION`(String) - (Required) Used for Read. Pendo Engage Rest API version you want to use. For example: 3.0.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.

- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.

Limitations

The following are limitations for the Pendo connector:

- Pagination is not supported in Pendo.
- Filtration is supported only by the Aggregate API objects(`Account`, `Event`, `Feature Event`, `Guide Events`, `Page Event`, `Poll Event`, `Track Event`, and `Visitor`)
- DateTimeRange is mandatory filter parameter for Aggregate API objects (`Event`, `Feature Event`, `Guide Events`, `Page Event`, `Poll Event, Track Event`)
- The dayRange period will be rounded down to the start of the period in the time zone. For example, if provided filter is `2023-01-12T07:55:27.065Z` then this time period will be rounded to the start of period, that is `2023-01-12T00:00:00Z` .

Connecting to Pipedrive

Pipedrive is a sales pipeline CRM designed to help small businesses manage leads, track sales activities and close more deals. Pipedrive enables sales teams in small businesses to: Streamline processes and consolidate sales data in one unified CRM sales tool. If you're a Pipedrive user, you can connect AWS Glue to your Pipedrive account. Then, you can use Pipedrive as a data source in your ETL Jobs. Run these jobs to transfer data between Pipedrive and AWS services or other supported applications.

Topics

- AWS Glue support for Pipedrive
- Policies containing the API operations for creating and using connections

- Configuring Pipedrive
- Configuring Pipedrive connections
- Reading from Pipedrive entities
- Pipedrive connection option reference
- Limitations

AWS Glue support for Pipedrive

AWS Glue supports Pipedrive as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Pipedrive.

Supported as a target?

No.

Supported Pipedrive API versions

v1.

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

You can also use the following managed IAM policies to allow access:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Pipedrive

Before you can use AWS Glue to transfer data from Pipedrive, you must meet these requirements:

Minimum requirements

- You have a Pipedrive account.
- Your Pipedrive account is enabled for API access.

If you meet these requirements, you’re ready to connect AWS Glue to your Pipedrive account. For typical connections, you don't need do anything else in Pipedrive.

Configuring Pipedrive connections

Pipedrive supports AUTHORIZATION_CODE grant type for OAuth2.

- This grant type is considered “three-legged” OAuth as it relies on redirecting users to the third party authorization server to authenticate the user. It is used when creating connections via the AWS Glue Console. The user creating a connection may by default rely on a AWS Glue-owned connected app where they do not need to provide any OAuth related information except for the

Pipedrive instanceurl. The AWS Glue Console will redirect the user to Pipedrive where the user must login and allow AWS Glue the requested permissions to access their Pipedrive instance.

- Users should opt to create their own connected app in Pipedrive and provide their own client id and client secret when creating connections through the AWS Glue Console. In this scenario, they will still be redirected to Pipedrive to login and authorize AWS Glue to access their resources.
- This grant type results in a refresh token and access token. The access token is active for one hour, and may be refreshed automatically without user interaction using the refresh token.
- For more information, see documentation on creating a connected app for AUTHORIZATION_CODE OAuth flow.

To configure a Pipedrive connection:

1. In AWS Secrets Manager, create a secret with the following details. It is required to create a secret for each connection in AWS Glue.
   a. For customer managed connected app - Secret should contain the connected app Consumer Secret with USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET as key.
2. In AWS Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. Under Data Connections, choose **Create connection**.
   b. When selecting a **Data Source**, select Pipedrive.
   c. Provide your Pipedrive **instanceURL**.
   d. Select the IAM role that AWS Glue can assume and has permissions for following actions:

      JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
```

```
        }
      ]
    }
```

e. Provide the User Managed Client Application ClientId of the Pipedrive you want to connect to.

f. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.

g. Select the network options if you want to use your network.

3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`. Choose **Next**.
4. Provide the **connectionName** and choose **Next**.
5. On the next page choose **Create connection**. You will be asked to login into Pipedrive. Provide your username and password and choose **Log In**.
6. Once you are logged in, choose **Continue to the App**. Now your connection is ready to be used.
7. In your AWS Glue job configuration, provide `connectionName` as an **Additional network connection**.

Reading from Pipedrive entities

Prerequisites

- A Pipedrive Object you would like to read from. Refer the supported entities table below to check the available entities.

Supported entities

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Activities	Yes	Yes	No	Yes	Yes
Activity Type	No	No	No	Yes	No
Call Logs	No	No	No	Yes	No
Currencies	Yes	Yes	No	Yes	No

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Deals	Yes	Yes	Yes	Yes	Yes
Leads	Yes	Yes	Yes	Yes	No
Lead Sources	No	Yes	No	Yes	No
Lead Labels	No	No	No	No	No
Notes	Yes	Yes	Yes	Yes	Yes
Organization	Yes	Yes	No	Yes	Yes
Permission Sets	Yes	No	No	Yes	No
Persons	Yes	Yes	Yes	Yes	Yes
Pipelines	No	Yes	No	Yes	No
Products	Yes	Yes	No	Yes	Yes
Roles	No	Yes	No	Yes	No
Stages	Yes	Yes	No	Yes	No
Users	No	No	No	Yes	No

Example

```
pipedrive_read= glueContext.create_dynamic_frame.from_options(
    connection_type="PIPEDRIVE",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "activites",
        "API_VERSION": "v1"
    }
```

Pipedrive entity and field details

Entities list:

- Activities: https://developers.pipedrive.com/docs/api/v1/Activities
- Activity Type: https://developers.pipedrive.com/docs/api/v1/ActivityTypes
- Call Logs: https://developers.pipedrive.com/docs/api/v1/CallLogs
- Currencies: https://developers.pipedrive.com/docs/api/v1/Currencies
- Deals: https://developers.pipedrive.com/docs/api/v1/Deals
- Leads: https://developers.pipedrive.com/docs/api/v1/Leads
- Lead Sources: https://developers.pipedrive.com/docs/api/v1/LeadSources
- Lead Labels: https://developers.pipedrive.com/docs/api/v1/LeadLabels
- Notes: https://developers.pipedrive.com/docs/api/v1/Notes
- Organizations: https://developers.pipedrive.com/docs/api/v1/Organizations
- Permission Sets: https://developers.pipedrive.com/docs/api/v1/PermissionSets
- Persons: https://developers.pipedrive.com/docs/api/v1/Persons
- Pipelines: https://developers.pipedrive.com/docs/api/v1/Pipelines
- Products: https://developers.pipedrive.com/docs/api/v1/Products
- Roles: https://developers.pipedrive.com/docs/api/v1/Roles
- Stages: https://developers.pipedrive.com/docs/api/v1/Stages
- Users: https://developers.pipedrive.com/docs/api/v1/Users

Entity	Data Type	Supported Operators
Activities, Deals, Notes, Organization, Persons and Products.	Date	'='
	Integer	'='
	String	'='
	Boolean	'='

Partitioning queries

In Pipedrive, only one field (due_date) from Activities entity supports field-based partitioning. It is a Date field.

Additional spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, `NUM_PARTITIONS` can be provided if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition query.
- `LOWER_BOUND`: an inclusive lower bound value of the chosen partition field.

  For date, we accept the Spark date format used in Spark SQL queries. Example of valid values: `"2024-02-06"`.
- `UPPER_BOUND`: an exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: number of partitions.

Example

```
pipedrive_read = glueContext.create_dynamic_frame.from_options(
    connection_type="PIPEDRIVE",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "activites",
        "API_VERSION": "v1",
        "PARTITION_FIELD": "due_date"
        "LOWER_BOUND": "2023-09-07T02:03:00.000Z"
        "UPPER_BOUND": "2024-05-07T02:03:00.000Z"
        "NUM_PARTITIONS": "10"
    }
```

Pipedrive connection option reference

The following are connection options for Pipedrive:

- `ENTITY_NAME`(String) - (Required) Used for Read/Write. The name of your Object in Pipedrive.
- `API_VERSION`(String) - (Required) Used for Read/Write. Pipedrive Rest API version you want to use. Example: v1.

- `INSTANCE_URL`(String) - (Required) URL of the instance where the user wants to run the operations. Example: v1.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.

Limitations

The following are limitations for the Pipedrive connector:

- Pipedrive supports field based partitioning for only one entity (Activities).
- Pipedrive supports record based partitioning for Activities, Deals, Notes, Persons, Organizations and Products entities.
- In Deals entity, status field as filter will returned all record if invalid value filter value is used.
- In Deals entity, ordering with multiple fields is not supported.
- To obtain performance data, we utilize a local AWS account. However, due to the limitation of refreshing the access token locally, the AWS Glue job for processing 1 GB of data is failing. Consequently, we have optimized the performance test with 179 MB of data, and the above results are based on this optimization. Nevertheless, we have observed that with an increasing number of partitions, the SaaS endpoint is taking more time compared to a single partition. We have consulted with the Pipedrive support team regarding this behavior, and they informed us that Pipedrive is silently throttling requests and delaying the response. Therefore, when running the AWS Glue job with large datasets or calling the same API endpoint multiple times, it may result in a timeout issue due to Pipedrive API's implementation. However, the connector and shim response times are decreasing as expected with an increase in the number of partitions.

Connecting to Productboard

Productboard is the product management system that helps product teams get the right products to market, faster. Over 3,000 modern, product-led companies, like Zendesk, UiPath, and Microsoft, use Productboard to understand what users really need, prioritize what to build next, and rally everyone around their roadmap.

Topics

- AWS Glue support for Productboard
- Policies containing the API operations for creating and using connections
- Configuring Productboard
- Configuring Productboard connections
- Reading from Productboard entities
- Productboard connection options
- Creating an Productboard account
- Limitations

AWS Glue support for Productboard

AWS Glue supports Productboard as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Productboard.

Supported as a target?

No.

Supported Productboard API versions

v1

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the preceding method, alternatively, use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Productboard

Before you can use AWS Glue to transfer from Productboard, you must meet the following requirements:

Minimum requirements

- You have an Productboard account with email and password. For more information about creating an account, see Creating a Productboard account.
- You must have AWS Account created with the service access to AWS Glue.
- You have a Productboard account's authentication details - either JWT Token if one want to use Custom Auth or Client ID and secret if one want to use OAuth2.0.
- If user wants to use `OAuth2.0`, Register your application with Productboard and setup the application by following the instructions at, How to integrate with Productboard via OAuth2 - developer documentation.

If you meet these requirements, you're ready to connect AWS Glue to your Productboard account. For typical connections, you don't need do anything else in Productboard.

Configuring Productboard connections

Productboard supports custom authentication and `OAuth2.0`. For `OAuth2.0` Productboard supports the `AUTHORIZATION_CODE` grant type.

- This grant type is considered "three-legged" `OAuth` as it relies on redirecting users to the third party authorization server to authenticate the user. It is used when creating connections via the AWS Glue Console. The user creating a connection may by default rely on a AWS Glue owned connected app where they do not need to provide any `OAuth` related information except for their Productboard Client ID and Client Secret. The AWS Glue Console will redirect the user to Productboard where the user must login and allow AWS Glue the requested permissions to access their Productboard instance.
- Users may still opt to create their own connected app in Productboard and provide their own Client ID and Client Secret when creating connections through the AWS Glue Console. In this scenario, they will still be redirected to Productboard to login and authorize AWS Glue to access their resources.
- This grant type results in a refresh token and access token. The access token is short lived, and may be refreshed automatically without user interaction using the refresh token.
- For public Productboard documentation on creating a connected app for `AUTHORIZATION_CODE` `OAuth` flow, see How to integrate with Productboard via OAuth2 - developer documentation.

To configure a Productboard connection:

1. In AWS Secrets Manager, create a secret with the following details:
   - For `OAuth` auth – For customer managed connected app: Secret should contain the connected app Consumer Secret with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as key.
   - For `Custom auth` – For customer managed connected app: Secret should contain the connected app `JWT token` with `access_token` as key.

   > **Note**
   >
   > It is a must to create a secret per connection in AWS Glue.

2. In AWS Glue Studio, create a connection under **Data Connections** by following the steps below:

   a. When selecting a **Data Source**, select Productboard.

   b. Select the IAM role for which AWS Glue can assume and has permissions for following actions:

   JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

   c. Select Authentication Type to connect to data source:

   - For `OAuth` auth – Provide the `Token URL`, and `User Managed Client Application ClientId` of the Productboard app.

   d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.
   e. Select the network options if you want to use your network.
3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.
4. In your AWS Glue job configuration, provide `connectionName` as an Additional network connection.

Reading from Productboard entities

Prerequisites

An Productboard Object you would like to read from. Refer the supported entities table below to check the available entities.

Supported entities

- Abuse-reports
- Automation
- Campaigns
- Click-details
- Lists
- Members
- Open-details
- Segments
- Stores
- Unsubscribed

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Features	Yes	Yes	No	Yes	Yes
Components	No	Yes	No	Yes	No
Products	No	Yes	No	Yes	No

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Feature Statuses	No	Yes	No	Yes	Yes
Custom Field Definitions	No	Yes	No	Yes	No
Custom Field Values	Yes	Yes	No	Yes	No

Example

```
Productboard_read = glueContext.create_dynamic_frame.from_options(
    connection_type="Productboard",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "feature",
        "API_VERSION": "1"
    }
```

Productboard entity and field details

- Features
- Components
- Feature statuses
- Products
- Custom fields definitions
- Custom fields values

Productboard connection options

The following are connection options for Productboard:

- `ENTITY_NAME`(String) – (Required) Used for Read/Write. The name of your Object in Productboard.

- `API_VERSION`(String) - (Required) Used for Read. Productboard Engage Rest API version you want to use. For example: 3.0.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.

Creating an Productboard account

1. Navigate to the Productboard sign up page, enter your email ID and password, and then choose **Log me in**.
2. In the **Account Name** field, enter the name of your Productboard account, and then select the **I agree to the Privacy Policy check** box.
3. On the **Now create your workspace** page, in the **Workspace URL** field, enter the URL for your new workspace. Then choose **Continue** to proceed to the next page and provide the remaining details.

   This creates your trial account. The trial account is free for 15 days. After the trial period expires, you can purchase a paid plan. Make a note of your email address, password, and workspace URL. You will need this information to access your account in the future."

Registering an `OAuth2.0` application

1. Navigate to the Productboard login page, enter your email ID and password, and choose **Log in**.
2. Select the **User** icon in the upper-right corner, and then choose **Account and billing** from the dropdown menu.
3. Select **Extras** and choose **Registered apps** from the dropdown menu.
4. Locate and choose **Register An App**.
5. Enter the following details:
   - **App name –** Name of the app.
   - **Company / Organization –** Name of your Company or Organization.
   - **App website –** Website of the app.

- **Redirect URI** – A Redirect URI pattern is a URI path (or comma-separated list of paths) to which Productboard can redirect (if requested) when the login flow is complete. For example, `https://ap-southeast-2\\.console\\.aws\\.amazon\\.com`

6. Choose **Create**.
7. The **Client ID** and **Client Secret** will now be visible. Copy and save them in a secure location. Then, choose **Done**.

> **Note**
>
> Your Client ID and Client Secret strings are credentials used to establish a connection with this connector when using AppFlow or AWS Glue.

Retrieving CustomAuth Credentials

1. Navigate to the Productboard login page, enter your email ID and password, and choose **Log me in**.

   You will be redirected to the home page.
2. On the home page, navigate to **Workspace Settings** > **Integrations** > **Public APIs** > **Access Token**.

> **Note**
>
> If the **Public APIs** section is not visible, your account might be on the Essentials plan. Access to API tokens requires at least a Pro plan. Plan features and names are subject to change. For more information about the packages, see Productboard pricing.

3. Choose **+** to generate a new token, and make sure to securely store it for future reference.

Creating `OAuth2.0` credentials

To utilize `OAuth2.0` authentication with the Productboard connector, you need to register your application on the Productboard platform and generate a `Client ID` and `Client Secret`.

1. Navigate to the Productboard login page, enter your email ID and password, and choose **Log me in**.

2. To register new OAuth2 application with your Productboard account, navigate to [Producboard]() page.
3. Complete the required fields and select the necessary scopes for each entity you wish to access.

> **Note**
>
> You have chosen the following four scopes, which are required for the six supported entities.

4. **Redirect URL** must have the following format: `https://ap-southeast-2\\.console\\.aws\\.amazon\\.com`

> **Note**
>
> The Appflow redirect URLs are subject to change. Once available, please update the redirect URLs for the AWS Glue platform.

5. The **Client ID** and **Client Secret** will now be visible. Copy and save them in a secure location.
6. You can set up and verify `OAuth2` by following the steps in the [How to Integrate with Productboard via OAuth2 developer]() documentation.

Limitations

The following are limitations for the Productboard connector:

- Productboard doesn't support either field based or record based partitioning.

Connecting to QuickBooks

QuickBooks is a leading accounting application for small and medium-sized businesses. QuickBooks accounting applications date back to the 1980s as one of the first products by Intuit, and accordingly was originally desktop software. Today, QuickBooks offers several accounting and business financial applications as both installable software and cloud-based SaaS software. As a QuickBooks user, you can connect AWS Glue to your QuickBooks account. Then, you can use QuickBooks as a data source in your ETL jobs. Run these jobs to transfer data between QuickBooks and AWS services or other supported applications.

Topics

- AWS Glue support for QuickBooks
- Policies containing the API operations for creating and using connections
- Configuring QuickBooks
- Configuring QuickBooks connections
- Reading from QuickBooks entities
- QuickBooks connection options
- Limitations and notes for QuickBooks connector

AWS Glue support for QuickBooks

AWS Glue supports QuickBooks as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from QuickBooks.

Supported as a target?

No.

Supported QuickBooks API versions

The following QuickBooks API versions are supported:

- v3

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring QuickBooks

Before you can use AWS Glue to transfer data from QuickBooks, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- You have a QuickBooks account.
- Your QuickBooks account is enabled for API access.

For more information, see the following topics in QuickBooks documentation:

- Create an Intuit Account
- Create and start developing your app

If you meet these requirements, you’re ready to connect AWS Glue to your QuickBooks account. For typical connections, you don't need do anything else in QuickBooks.

Configuring QuickBooks connections

QuickBooks supports the AUTHORIZATION_CODE grant type for OAuth2. The grant type determines how AWS Glue communicates with QuickBooks to request access to your data.

- This grant type is considered "three-legged" OAuth as it relies on redirecting users to a third-party authorization server to authenticate the user. It is used when creating connections via the AWS Glue console.
- Users may still opt to create their own connected app in QuickBooks and provide their own client ID and client secret when creating connections through the AWS Glue console. In this scenario, they will still be redirected to QuickBooks to login and authorize AWS Glue to access their resources.
- This grant type results in a refresh token and access token. The access token is short lived, and may be refreshed automatically without user interaction using the refresh token.
- For public QuickBooks documentation on creating a connected app for Authorization Code OAuth flow, see Set up OAuth 2.0.

To configure a QuickBooks connection:

1. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Connection type**, select QuickBooks.
   b. Provide the instance URL and company ID of the QuickBooks instance you want to connect to.
   c. Select the AWS IAM role which AWS Glue can assume and has permissions for following actions:

      JSON

      ```
      {
        "Version": "2012-10-17",
        "Statement": [
      ```

```
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

   d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.
   e. Select the network options if you want to use your network.

2. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from QuickBooks entities

Prerequisite

A QuickBooks object you would like to read from.

Supported entities for source:

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Account	Yes	Yes	Yes	Yes	Yes
Bill	Yes	Yes	Yes	Yes	Yes
Company Info	No	No	No	Yes	No
Customer	Yes	Yes	Yes	Yes	Yes
Employee	Yes	Yes	Yes	Yes	Yes

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Estimate	Yes	Yes	Yes	Yes	Yes
Invoice	Yes	Yes	Yes	Yes	Yes
Item	Yes	Yes	Yes	Yes	Yes
Payment	Yes	Yes	Yes	Yes	Yes
Preferences	No	No	No	Yes	No
Profit and Loss	Yes	No	No	Yes	No
Tax Agency	Yes	Yes	Yes	Yes	Yes
Vendors	Yes	Yes	Yes	Yes	Yes

Example:

```
QuickBooks_read = glueContext.create_dynamic_frame.from_options(
    connection_type="quickbooks",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "Account",
        "API_VERSION": "v3"
    }
```

QuickBooks entity and field details:

For more information about the entities and field details see:

- Account
- Bill
- CompanyInfo
- Customer
- Employee

- Estimate
- Invoice
- Item
- Payment
- Preferences
- ProfitAndLoss
- TaxAgency
- Vendor

Partitioning queries

Field-based partitioning:

In QuickBooks, the Integer and DateTime datatype fields support field-based partitioning.

You can provide the additional Spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, and `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition the query.
- `LOWER_BOUND`: an **inclusive** lower bound value of the chosen partition field.

  For the Datetime field, we accept the Spark timestamp format used in Spark SQL queries.

  Examples of valid value:

  ```
  "2024-05-07T02:03:00.00Z"
  ```
- `UPPER_BOUND`: an **exclusive** upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: the number of partitions.

Example:

```
QuickBooks_read = glueContext.create_dynamic_frame.from_options(
    connection_type="quickbooks",
    connection_options={
```

```
        "connectionName": "connectionName",
        "REALMID": "12345678690123456789",
        "ENTITY_NAME": "Account",
        "API_VERSION": "v3",
        "PARTITION_FIELD": "MetaData_CreateTime"
        "LOWER_BOUND": "2023-09-07T02:03:00.000Z"
        "UPPER_BOUND": "2024-05-07T02:03:00.000Z"
        "NUM_PARTITIONS": "10"
    }
```

Record-based partitioning:

The original query is splitinto `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently:

- `NUM_PARTITIONS`: the number of partitions.

Example:

```
QuickBooks_read = glueContext.create_dynamic_frame.from_options(
    connection_type="quickbooks",
    connection_options={
        "connectionName": "connectionName",
        "REALMID": "1234567890123456789",
        "ENTITY_NAME": "Bill",
        "API_VERSION": "v3",
        "NUM_PARTITIONS": "10"
    }
```

QuickBooks connection options

The following are connection options for QuickBooks:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your object in QuickBooks.
- `INSTANCE_URL`(String) - (Required) A valid QuickBooks instance URL.
- `API_VERSION`(String) - (Required) Used for Read. QuickBooks Rest API version you want to use.
- `REALM_ID`(String) - An ID that identifies an individual QuickBooks Online company where you send requests.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.

- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.

Limitations and notes for QuickBooks connector

The following are limitations or notes for the QuickBooks connector:

- In the `taxAgency` API, order by filtering is not working as expected.

Connecting to Salesforce

Salesforce provides customer relationship management (CRM) software that help you with sales, customer service, e-commerce, and more. If you're a Salesforce user, you can connect AWS Glue to your Salesforce account. Then, you can use Salesforce as a data source or destination in your ETL Jobs. Run these jobs to transfer data between Salesforce and AWS services or other supported applications.

Topics

- AWS Glue support for Salesforce
- Policies containing the API operations for creating and using connections
- Configuring Salesforce
- Apply System Admin profile
- Configuring Salesforce connections
- Reading from Salesforce
- Writing to Salesforce
- Salesforce connection options
- Limitations for the Salesforce connector

- Set up the Authorization Code flow for Salesforce
- Set up the JWT bearer OAuth flow for Salesforce

AWS Glue support for Salesforce

AWS Glue supports Salesforce as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Salesforce.

Supported as a target?

Yes. You can use AWS Glue ETL jobs to write records into Salesforce.

Supported Salesforce API versions

The following Salesforce API versions are supported

- v58.0
- v59.0
- v60.0

Policies containing the API operations for creating and using connections

The following sample IAM policy describes the required permissions for creating, managing and using Salesforce connections within AWS Glue ETL jobs. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

You can also use the following IAM policies to allow access:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

If providing Network Options when creating a Salesforce connection, the following actions must also be included in the IAM role:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

For Zero-ETL Salesforce connections, see Zero-ETL prerequisites.

For Zero-ETL Salesforce connections, see Zero-ETL prerequisites.

Configuring Salesforce

Before you can use AWS Glue to transfer data to or from Salesforce, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- You have a Salesforce account.
- Your Salesforce account is enabled for API access. API access is enabled by default for the Enterprise, Unlimited, Developer, and Performance editions.

If you meet these requirements, you’re ready to connect AWS Glue to your Salesforce account. AWS Glue handles the remaining requirements with the AWS managed connected app.

The AWS managed connected app for Salesforce

The AWS managed connected app helps you create a Salesforce connection in fewer steps. In Salesforce, a connected app is a framework that authorizes external applications, like AWS Glue, to access your Salesforce data using OAuth 2.0. To use the AWS managed connected app, create a Salesforce connection by using the AWS Glue consule. When you configure the connection, set the **OAuth grant type** to **Authorization code** and leave the box checked for **Use AWS managed client application**.

When saving the connection, you will be redirected to Salesforce to login and approve AWS Glue access to your Salesforce account.

Apply System Admin profile

In Salesforce, follow the steps to apply the System Admin profile:

1. In Salesforce, navigate to **Settings > Connected Apps > Connected Apps OAuth Usage**.
2. In the list of connected apps, find AWS Glue and choose **Install**. If needed, choose **Unblock**.
3. Navigate to **Settings > Manage Connected Apps then choose AWS Glue**. Under OAuth Policies, choose **Admin approved users are pre-authorized** and select the **System Admin** profile. This action restricts access to AWS Glue only to users with the System Admin profile.

Apply System Admin profile

In Salesforce, follow the steps to apply the System Admin profile:

1. In Salesforce, navigate to **Settings > Connected Apps > Connected Apps OAuth Usage**.
2. In the list of connected apps, find AWS Glue and choose **Install**. If needed, choose **Unblock**.
3. Navigate to **Settings > Manage Connected Apps then choose AWS Glue**. Under OAuth Policies, choose **Admin approved users are pre-authorized** and select the **System Admin** profile. This action restricts access to AWS Glue only to users with the System Admin profile.

Configuring Salesforce connections

To configure a Salesforce connection:

1. In AWS Secrets Manager, create a secret with the following details:
   a. For the JWT_TOKEN grant type - the secret should contain the JWT_TOKEN key with its value.
   b. For the AuthorizationCode grant type:
      i. For an AWS Managed connected app, an empty secret or a secret with some temporary value must be provided.
      ii. For a customer managed connected app, the secret should contain the connected app `Consumer Secret` with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as the key.
   c. Note: You must create a secret for your connection in AWS Glue.
2. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Connection type**, select Salesforce.
   b. Provide the INSTANCE_URL of the Salesforce instance you want to connect to.
   c. Provide the Salesforce environment.

d. Select the AWS IAM role which AWS Glue can assume and has permissions for following actions:

   JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

e. Select the OAuth2 grant type which you want to use for the connections. The grant type determines how AWS Glue communicates with Salesforce to request access to your data. Your choice affects the requirements that you must meet before you create the connection. You can choose either of these types:

   - **JWT_BEARER Grant Type**: This grant type works well for automation scenarios as it allows a JSON Web Token (JWT) to be created up front with the permissions of a particular user in the Salesforce instance. The creator has control over how long the JWT is valid for. AWS Glue is able to use the JWT to obtain an access token which is used to call Salesforce APIs.

     This flow requires that the user has created a connected app in their Salesforce instance which enables issuing JWT-based access tokens for users.

     For information on creating a connected app for the JWT bearer OAuth flow, see OAuth 2.0 JWT bearer flow for server-to-server integration. To set up the JWT bearer flow with the Salesforce connected app, see Set up the JWT bearer OAuth flow for Salesforce.
   - **AUTHORIZATION_CODE Grant Type**: This grant type is considered a "three-legged" OAuth as it relies on redirecting users to the third-party authorization server to authenticate the

user. It is used when creating connections via the AWS Glue console. The user creating a connection may by default rely on an AWS Glue connected app (AWS Glue managed client application) where they do not need to provide any OAuth related information except for their Salesforce instance URL. The AWS Glue console will redirect the user to Salesforce where the user must login and allow AWS Glue the requested permissions to access their Salesforce instance.

Users may still opt to create their own connected app in Salesforce and provide their own client ID and client secret when creating connections through the AWS Glue console. In this scenario, they will still be redirected to Salesforce to login and authorize AWS Glue to access their resources.

This grant type results in a refresh token and access token. The access token is short lived, and may be refreshed automatically without user interaction using the refresh token.

For information on creating a connected app for the Authorization Code OAuth flow, see the section called “Set up the Authorization Code flow for Salesforce”.

   f. Select the `secretName` which you want to use for this connection in AWS Glue to store the OAuth 2.0 tokens.
   g. Select the network options if you want to use your network.

3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.
4. If providing network options, also grant the IAM role the following permissions:

   JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

Configuring Salesforce connections with the AWS CLI

You can create Salesforce connections using the AWS CLI:

```
aws glue create-connection --connection-input \
"{\"Name\": \"salesforce-conn1\",\"ConnectionType\": \"SALESFORCE\",
\"ConnectionProperties\": {\"ROLE_ARN\": \"arn:aws:iam::123456789012:role/glue-role\",
\"INSTANCE_URL\": \"https://example.my.salesforce.com\"},\"ValidateCredentials\": true,
\"AuthenticationConfiguration\": {\"AuthenticationType\": \"OAUTH2\",\"SecretArn\":
 \"arn:aws:secretsmanager:us-east-1:123456789012:secret:salesforce-conn1-secret-IAmcdk
\",\"OAuth2Properties\": {\"OAuth2GrantType\": \"JWT_BEARER\",\"TokenUrl\": \"https://
login.salesforce.com/services/oauth2/token\"}}}" \
--endpoint-url https://glue.us-east-1.amazonaws.com \
--region us-east-1
```

Reading from Salesforce

Prerequisite

A Salesforce sObject you would like to read from. You will need the object name such as `Account` or `Case` or `Opportunity`.

Example:

```
salesforce_read = glueContext.create_dynamic_frame.from_options(
    connection_type="salesforce",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "Account",
        "API_VERSION": "v60.0"
    }
```

Partitioning queries

You can provide the additional Spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, and `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition the query.
- `LOWER_BOUND`: an **inclusive** lower bound value of the chosen partition field.

For Date or Timestamp fields, the connector accepts the Spark timestamp format used in Spark SQL queries.

Examples of valid values:

```
"TIMESTAMP \"1707256978123\""
"TIMESTAMP '2018-01-01 00:00:00.000 UTC'"
"TIMESTAMP \"2018-01-01 00:00:00 Pacific/Tahiti\""
"TIMESTAMP \"2018-01-01 00:00:00\""
"TIMESTAMP \"-123456789\" Pacific/Tahiti"
"TIMESTAMP \"1702600882\""
```

- `UPPER_BOUND`: an **exclusive** upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: the number of partitions.
- `TRANSFER_MODE`: supports two modes: `SYNC` and `ASYNC`. Default is `SYNC`. When set to `ASYNC`, Bulk API 2.0 Query will be utilized for processing.

Example:

```
salesforce_read = glueContext.create_dynamic_frame.from_options(
    connection_type="salesforce",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "Account",
        "API_VERSION": "v60.0",
        "PARTITION_FIELD": "SystemModstamp",
        "LOWER_BOUND": "TIMESTAMP '2021-01-01 00:00:00 Pacific/Tahiti'",
        "UPPER_BOUND": "TIMESTAMP '2023-01-10 00:00:00 Pacific/Tahiti'",
        "NUM_PARTITIONS": "10",
        "TRANSFER_MODE": "ASYNC"
    }
```

Writing to Salesforce

Prerequisites

A Salesforce sObject you would like to write to. You will need the object name such as `Account` or `Case` or `Opportunity`.

The Salesforce connector supports four write operations:

- INSERT
- UPSERT
- UPDATE
- DELETE

When using the `UPSERT` write operation, the `ID_FIELD_NAMES` option must be provided to specify the external ID field for the records.

You can also add connection options:

- `TRANSFER_MODE`: Supports two modes: `SYNC` and `ASYNC`. Default is `SYNC`. When set to `ASYNC`, Bulk API 2.0 Ingest will be utilized for processing.
- `FAIL_ON_FIRST_ERROR`: The default value is `FALSE`, which means the AWS Glue job will continue processing all the data even if there are some failed write records. When set to `TRUE`, the AWS Glue job will fail if there are any failed write records, and it will not continue processing.

Example

```
salesforce_write = glueContext.write_dynamic_frame.from_options(
    frame=frameToWrite,
    connection_type="salesforce",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "Account",
        "API_VERSION": "v60.0",
        "WRITE_OPERATION": "INSERT",
        "TRANSFER_MODE": "ASYNC",
        "FAIL_ON_FIRST_ERROR":""true"
    }
```

Salesforce connection options

The following connection options are supported for the Salesforce connector:

- `ENTITY_NAME`(String) - (Required) Used for Read/Write. The name of your Object in Salesforce.
- `API_VERSION`(String) - (Required) Used for Read/Write. Salesforce Rest API version you want to use.

- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.

  When providing a filter predicate, only the `AND` operator is supported. Other operators such as `OR` and `IN` are not currently supported.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.
- `IMPORT_DELETED_RECORDS`(String) - Default: FALSE. Used for read. To get the delete records while querying.
- `WRITE_OPERATION`(String) - Default: INSERT. Used for write. Value should be INSERT, UPDATE, UPSERT, DELETE.
- `ID_FIELD_NAMES`(String) - Default : null. Required for UPDATE and UPSERT.

Limitations for the Salesforce connector

The following are limitations for the Salesforce connector:

- We only support Spark SQL and Salesforce SOQL is not supported.
- Job bookmarks are not supported.
- Salesforce field names are case sensitive. When writing to Salesforce, data must match the casing of the fields defined within Salesforce.

Set up the Authorization Code flow for Salesforce

Refer to Salesforce public documentation for enabling the OAuth 2.0 Authorization Code flow.

To configure the connected app:

1. Activate the **Enable OAuth Settings** checkbox.
2. In the **Callback URL** text field, enter one or more redirect URLs for AWS Glue.

   Redirect URLs have the following format:

   https://`region`.console.aws.amazon.com/gluestudio/oauth

   In this URL, *region* is the code for the AWS Region where you use AWS Glue to transfer data from Salesforce. For example, the code for the US East (N. Virginia) Region is `us-east-1`. For that Region, the URL is the following:

   https://us-east-1.console.aws.amazon.com/gluestudio/oauth

   For the AWS Regions that AWS Glue supports, and their codes, see AWS Glue endpoints and quotas in the *AWS General Reference*.
3. Activate the **Require Secret for Web Server Flow** checkbox.
4. In the **Available OAuth Scopes** list, add the following scopes:
   - Manage user data via APIs (api)
   - Access custom permissions (custom_permissions)
   - Access the identity URL service (id, profile, email, address, phone)
   - Access unique user identifiers (openid)
   - Perform requests at any time (refresh_token, offline_access)
5. Set the refresh token policy for the connected app to **Refresh token is valid until revoked**. Otherwise, your jobs will fail when your refresh token expires. For more information on how to check and edit the refresh token policy, see Manage OAuth Access Policies for a Connected App in the Salesforce documentation.

Set up the JWT bearer OAuth flow for Salesforce

Refer to Salesforce public documentation for enabling server-to-server integration with OAuth 2.0 JSON Web Tokens.

Once you have created a JWT and configured the connected app appropriately in Salesforce, you can create a new Salesforce connection with the `JWT_TOKEN` key set in your Secrets Manager Secret. Set the OAuth grant type to **JWT Bearer Token** when creating the connection.

Connecting to Salesforce Marketing Cloud

Salesforce Marketing Cloud is a provider of marketing automation and analytics software for email, mobile, social, and online marketing. It also offers consulting and implementation services. As a Salesforce Marketing Cloud user, you can connect AWS Glue to your Salesforce Marketing Cloud account. Then, you can use Salesforce Marketing Cloud as a data source or destination in your ETL jobs. Run these jobs to transfer data between Salesforce Marketing Cloud and AWS services or other supported applications.

Topics

- AWS Glue support for Salesforce Marketing Cloud
- Policies containing the API operations for creating and using connections
- Configuring Salesforce Marketing Cloud
- Configuring Salesforce Marketing Cloud connections
- Reading from Salesforce Marketing Cloud entities
- Salesforce Marketing Cloud connection options
- Limitations and notes for Salesforce Marketing Cloud connector

AWS Glue support for Salesforce Marketing Cloud

AWS Glue supports Salesforce Marketing Cloud as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Salesforce Marketing Cloud.

Supported as a target?

No.

Supported Salesforce Marketing Cloud API versions

The following Salesforce Marketing Cloud API versions are supported:

- v1

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Salesforce Marketing Cloud

Before you can use AWS Glue to transfer data from Salesforce Marketing Cloud, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- You have a Salesforce Marketing Cloud account. For more information, see Creating a Salesforce Marketing Cloud account.
- Your Salesforce Marketing Cloud account is enabled for API access. API access is enabled by default for the Enterprise, Unlimited, Developer, and Performance editions.

If you meet these requirements, you’re ready to connect AWS Glue to your Salesforce Marketing Cloud account. For typical connections, you don't need do anything else in Salesforce Marketing Cloud.

Creating a Salesforce Marketing Cloud account

For Salesforce Marketing cloud, you need to contact the vendor for account creation. If you or your company has an association with Salesforce, contact your Salesforce account manager to request a Salesforce Marketing Cloud license. Otherwise, you can request contact from a Salesforce representative as follows:

1. Go to https://www.salesforce.com/in/products/marketing-cloud/overview/ and choose **Sign up**.
2. Select the **Contact Us** link on the top right of the page.
3. Enter the required information in the form and choose **Contact Me**.

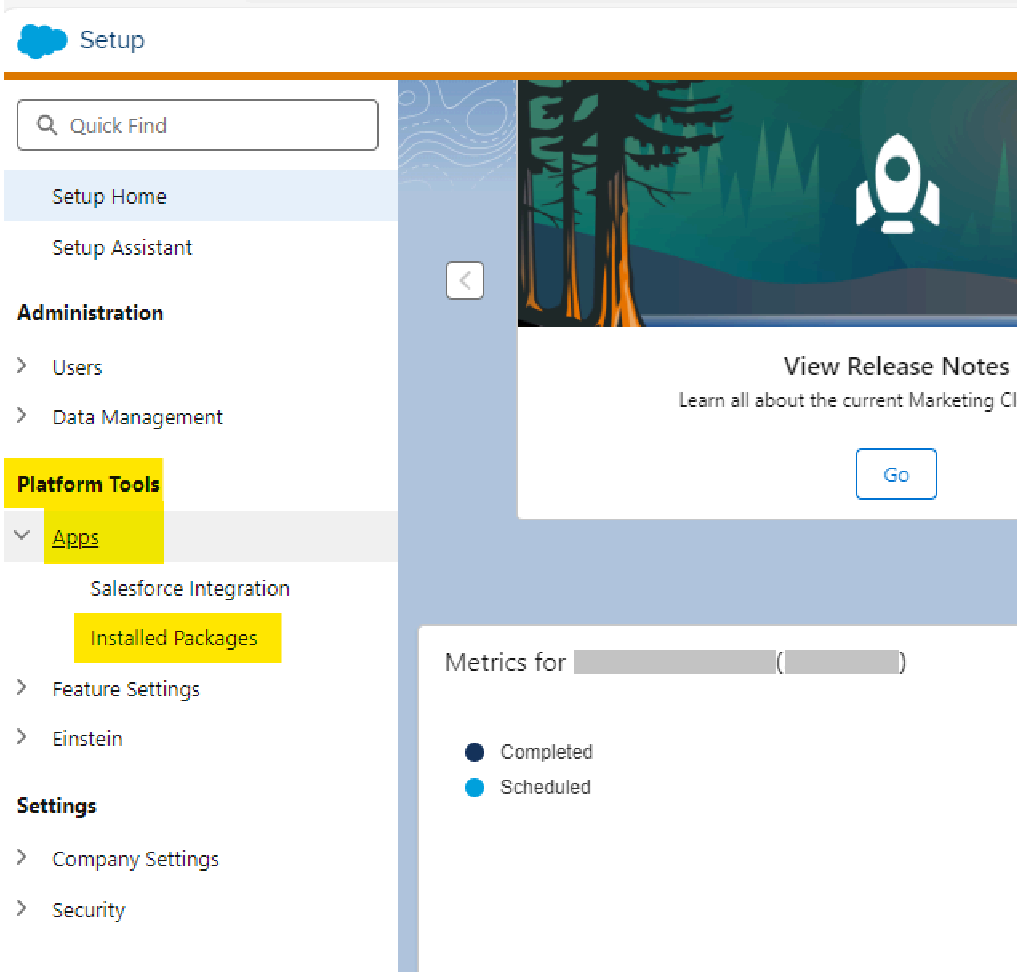
A Salesforce representative will contact you to discuss your requirements.

Creating a project and OAuth 2.0 credentials

To get a project and OAuth 2.0 credentials:

1. Log into your Salesforce Marketing Cloud instance with your username and password and authenticate using your registered mobile number.
2. Click on your profile at the top right corner and then go to **Setup**.

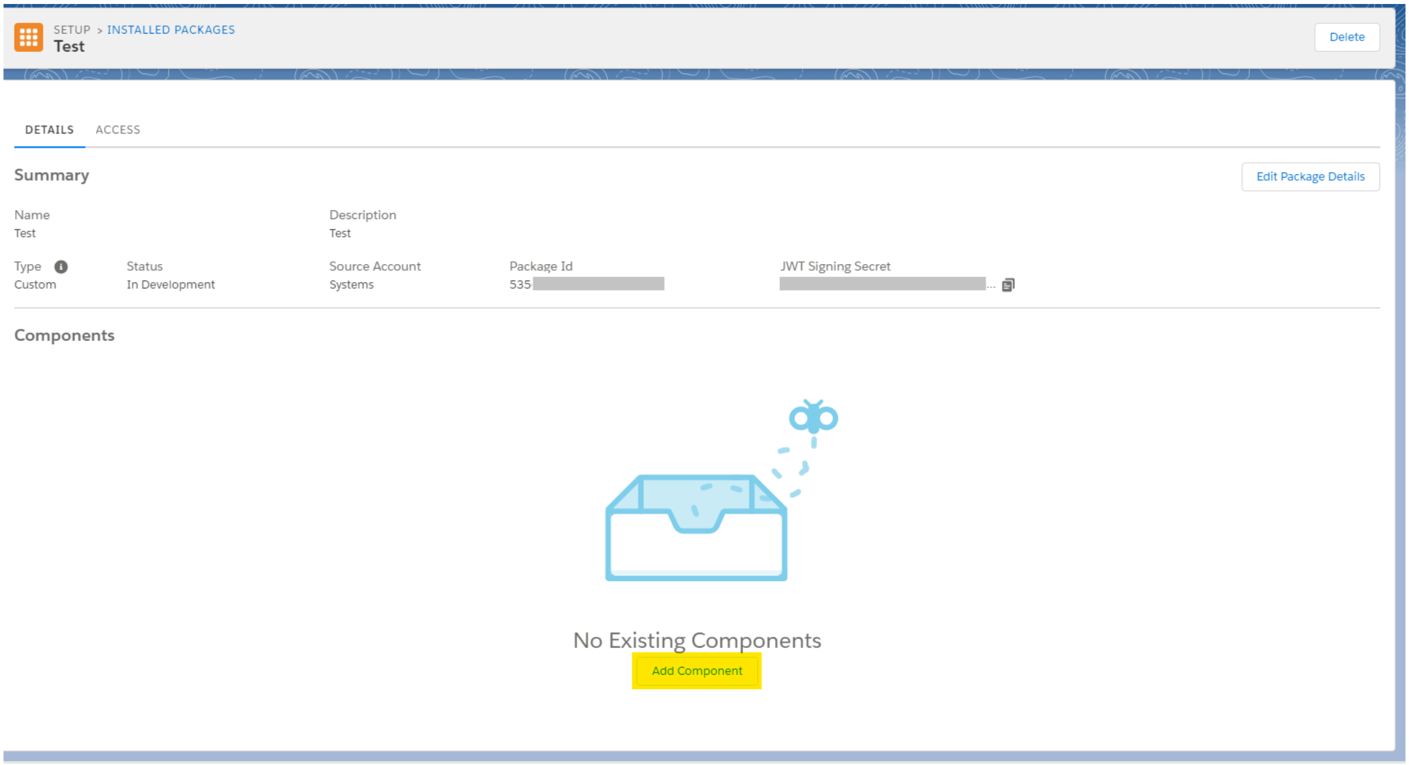
3. Under **Platform Tools** choose **Apps** and then choose **Installed Packages**.


Setup
Quick Find
Setup Home
Setup Assistant
Administration
Users
Data Management
Platform Tools
Apps
Salesforce Integration
Installed Packages
Feature Settings
Einstein
Settings
Company Settings
Security
View Release Notes
Learn all about the current Marketing Cl
Go
Metrics for [redacted] ([redacted])
Completed
Scheduled

4. On the **Installed Packages** page, click **New** at the top right corner. Provide the name and description of the package.

   Save the package. After the package is saved, you can view the package details.

5. On the **Details** page of the package, under the **Component** section, choose **Add Component**.



Test

Summary

Name
Test

Description
Test

Type
Custom

Status
In Development

Source Account
Systems

Package Id
535[redacted]

JWT Signing Secret
[redacted]...

Components

No Existing Components

6. Select the **Component Type** as 'API Integration' and click **Next**.
7. Select the **Integration Type** as 'Server-to-Server' (which has the client credentials OAuth grant type) and click **Next**.
8. Add the scopes based on your requirements and click **Save**.

Configuring Salesforce Marketing Cloud connections

Salesforce Marketing Cloud supports the CLIENT CREDENTIALS grant type for OAuth2.

- This grant type is considered 2-legged OAuth 2.0 as it is used by clients to obtain an access token outside of the context of a user. AWS Glue is able to use the client ID and client secret to authenticate the Salesforce Marketing Cloud APIs which are provided by custom services that you define.
- Each custom service is owned by an API-only user which has a set of roles and permissions which authorize the service to perform specific actions. An access token is associated with a single custom service.
- This grant type results in an access token which is short lived, and may be renewed by calling an identity endpoint.

- For public Salesforce Marketing Cloud documentation for OAuth 2.0 with client credentials, see Set Up Your Development Environment for Enhanced Packages.

To configure a Salesforce Marketing Cloud connection:

1. In AWS Secrets Manager, create a secret with the following details:
    a. For the customer managed connected app, the Secret should contain the connected app Consumer Secret with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as key.
    b. Note: You must create a secret for your connections in AWS Glue.
2. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
    a. When selecting a **Connection type**, select Salesforce Marketing Cloud.
    b. Provide the `Subdomain Endpoint` of the Salesforce Marketing Cloud you want to connect to.
    c. Select the AWS IAM role which AWS Glue can assume and has permissions for following actions:

       JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

    d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.

    e. Select the network options if you want to use your network.

3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Salesforce Marketing Cloud entities

Prerequisite

A Salesforce Marketing Cloud object you would like to read from. You will need the object name such as `Activity` or `Campaigns`. The following table shows the supported entities.

Supported entities for source:

Entity	Interface	Can be filtered	Supports limit	Supports Order by	Supports SELECT *	Supports partitioning
Event Notification Callback	REST	No	No	No	Yes	No
Seed-List	REST	No	Yes	No	Yes	No
Setup	REST	Yes	Yes	No	Yes	No
Domain Verification	REST	Yes	Yes	Yes	Yes	No
Objects Nested Tags	REST	Yes	No	No	Yes	No
Contact	REST	No	Yes	No	Yes	No
Event Notification	REST	No	No	No	Yes	No

Entity	Interface	Can be filtered	Supports limit	Supports Order by	Supports SELECT *	Supports partitioning
Subscription						
Messaging	REST	No	Yes	No	Yes	No
Activity	SOAP	No	No	No	Yes	Yes
Bounce Event	SOAP	No	No	No	Yes	Yes
Click Event	SOAP	No	No	No	Yes	Yes
Content Area	SOAP	No	No	No	Yes	Yes
Data Extension	SOAP	No	Yes	No	Yes	Yes
Email	SOAP	No	Yes	No	Yes	Yes
Forwarded Email Event	SOAP	No	Yes	No	Yes	Yes
Forward Email OptInEvent	SOAP	No	Yes	No	Yes	Yes
Link	SOAP	No	Yes	No	Yes	Yes
Link Send	SOAP	No	Yes	No	Yes	Yes
List	SOAP	No	Yes	No	Yes	Yes

Entity	Interface	Can be filtered	Supports limit	Supports Order by	Supports SELECT *	Supports partitioning
List Subscriber	SOAP	No	Yes	No	Yes	Yes
Not Sent Event	SOAP	No	Yes	No	Yes	Yes
Open Event	SOAP	No	Yes	No	Yes	Yes
Send	SOAP	No	Yes	No	Yes	Yes
Sent Event	SOAP	No	Yes	No	Yes	Yes
Subscriber	SOAP	No	Yes	No	Yes	Yes
Survey Event	SOAP	No	Yes	No	Yes	Yes
Unsub Event	SOAP	No	Yes	No	Yes	Yes
Audit Events	REST	No	Yes	Yes	Yes	No
Campaigns	REST	No	Yes	Yes	Yes	No
Interactions	REST	No	Yes	Yes	Yes	No
Content Assets	REST	No	Yes	Yes	Yes	No

Example for REST:

```
salesforcemarketingcloud _read = glueContext.create_dynamic_frame.from_options(
    connection_type="salesforcemarketingcloud",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "Campaigns",
        "API_VERSION": "v1",
        "INSTANCE_URL": "https://********************.rest.marketingcloudapis.com"
    }
```

Example for SOAP:

```
salesforcemarketingcloud _read = glueContext.create_dynamic_frame.from_options(
    connection_type="salesforcemarketingcloud",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "Activity",
        "API_VERSION": "v1",
        "INSTANCE_URL": "https://********************.soap.marketingcloudapis.com"
    }
```

Salesforce Marketing Cloud entity and field details:

The following tables describe the Salesforce Marketing Cloud entities. There are REST entities with static metadata and SOAP entities with dynamic metadata.

REST entities with static metadata:

Entity	Field	Data type	Supported operators
Event Notification Callback	callbackId	String	
	callbackName	String	
	url	String	
	maxBatchSize	Integer	
	status	String	
	statusReason	String	
Seed-List	id	String	
	name	String	

Entity	Field	Data type	Supported operators
	description	String	
	activeSeedCount	Integer	
Setup	customerKey	String	
	name	String	
	description	String	
	locationType	String	'='
	awsFileTransferLocation	Struct	
Domain verification	enterpriseId	Integer	
	status	String	'='
	domainType	String	'='
	memberId	Integer	
	emailSendTime	DateTime	
	domain	String	
	isSendable	Boolean	
Objects Nested Tags	id	Integer	
	modifiedDate	DateTime	
	tags	List	
	name	String	
	description	String	
	parentId	Integer	

Entity	Field	Data type	Supported operators
Contact	values	List	
Event Notification Subscription	subscriptionName	String	
	callbackId	String	
	callbackName	String	
	eventCategoryTypes	List	
	filters	List	
	url	String	
	maxBatchSize	Integer	
	subscriptionId	String	
	status	String	
	statusReason	String	
Messaging	deliveryTime	DateTime	
	id	String	
	messageId	String	
	status	String	
	to	Struct	
Interactions	status	String	'='
	id	String	
	key	String	
	name	String	

Entity	Field	Data type	Supported operators
	lastPublishedDate	DateTime	
	description	String	
	version	Integer	
	workflowApiVersion	Integer	
	createdDate	DateTime	
	modifiedDate	DateTime	
	goals	Struct	
	stats	Struct	
	entryMode	String	
	defaults	Struct	
	executionMode	Struct	
	definitionId	String	
Content Assets	id	Integer	
	customerKey	String	
	objectId	String	
	contentType	String	
	assetType	Struct	
	name	String	
	description	String	
	owner	Struct	

Entity	Field	Data type	Supported operators
	createdDate	DateTime	
	createdBy	Struct	
	modifiedDate	DateTime	
	modifiedBy	Struct	
	thumbnail	Struct	
	category	Struct	
	meta	Struct	
	views	Struct	
	availableViews	Struct	
	data	Struct	
	legacyData	Struct	
	modelVersion	Integer	
	Version	Integer	
	Locked	Boolean	
	FileProperties	Struct	
	Tags	List	
	Content	String	
	Design	String	
	SuperContent	String	
	CustomFields	Struct	

Entity	Field	Data type	Supported operators
	Blocks	Struct	
	MinBlocks	Integer	
	MaxBlocks	Integer	
	Channels	Struct	
	AllowedBlocks	List	
	Slots	Struct	
	BusinessUnitAvailability	Struct	
	sharingProperties	Struct	
	sharingProperties.sharedWith	Struct	
	sharingProperties.sharingType	String	
	Template	Struct	
	File	String	
	GenerateFrom	String	
Audit Events	id	Integer	
	createdDate	DateTime	
	memberId	Integer	
	enterpriseId	Integer	
	employee	Struct	
	objectType	Struct	

Entity	Field	Data type	Supported operators
	operation	Struct	
	object	Struct	
	transactionId	String	
Campaigns	id	Integer	
	createdDate	DateTime	
	modifiedDate	DateTime	
	name	String	
	description	String	
	campaignCode	String	
	color	String	
	favorite	Boolean	

SOAP entities with dynamic metadata:

Entity	Data type	Supported operators
Activity	String	LIKE,!=,=
	Struct	
	Integer	!=,=,>=,<=,<,>
	Double	!=,=,>=,<=,<,>
	Boolean	!=,=
	DateTime	>=,<=,<,>,=,BETWEEN

Entity	Data type	Supported operators
Bounce Event	Integer	!=,=,>=,<=,<,>
	DateTime	>=,<=,<,>,=,BETWEEN
	String	LIKE,!=,=
	Struct	
Click Event	Integer	!=,=,>=,<=,<,>
	DateTime	>=,<=,<,>,=,BETWEEN
	String	LIKE,!=,=
	Struct	
Content Area	Struct	
	String	LIKE,!=,=
	Integer	!=,=,>=,<=,<,>
	DateTime	>=,<=,<,>,=,BETWEEN
	Boolean	!=,=
Data Extension	DateTime	>=,<=,<,>,=,BETWEEN
	String	LIKE,!=,=
Email	Integer	!=,=,>=,<=,<,>
	String	LIKE,!=,=
	DateTime	>=,<=,<,>,=,BETWEEN
	Boolean	!=,=
	Struct	

Entity	Data type	Supported operators
Forwarded Email Event	Integer	!=,=,>=,<=,<,>
	String	LIKE,!=,=
	DateTime	>=,<=,<,>,=,BETWEEN
	Struct	
Forwarded Email OptInEvent	Integer	!=,=,>=,<=,<,>
	String	LIKE,!=,=
	DateTime	>=,<=,<,>,=,BETWEEN
	Struct	
Link	Integer	!=,=,>=,<=,<,>
Link Send	Integer	!=,=,>=,<=,<,>
	String	LIKE,!=,=
	Double	!=,=,>=,<=,<,>
List	Integer	!=,=,>=,<=,<,>
	String	LIKE,!=,=
	DateTime	>=,<=,<,>,=,BETWEEN
	Struct	
List Subscriber	Integer	!=,=,>=,<=,<,>
	String	LIKE,!=,=
	DateTime	>=,<=,<,>,=,BETWEEN
	Struct	

Entity	Data type	Supported operators
Not Sent Event	Integer	!=,=,>=,<=,<,>
	String	LIKE,!=,=
	DateTime	>=,<=,<,>,=,BETWEEN
	Struct	
Open Event	Integer	!=,=,>=,<=,<,>
	String	LIKE,!=,=
	DateTime	>=,<=,<,>,=,BETWEEN
	Struct	
Send	Integer	!=,=,>=,<=,<,>
	String	LIKE,!=,=
	DateTime	>=,<=,<,>,=,BETWEEN
	Boolean	!=,=
	Struct	
Sent Event	Integer	!=,=,>=,<=,<,>
	String	LIKE,!=,=
	DateTime	>=,<=,<,>,=,BETWEEN
	Struct	
Subscriber	Integer	!=,=,>=,<=,<,>
	String	LIKE,!=,=
	DateTime	>=,<=,<,>,=,BETWEEN

Entity	Data type	Supported operators
	Struct	
Survey Event	Integer	!=,=,>=,<=,<,>
	String	LIKE,!=,=
	DateTime	>=,<=,<,>,=,BETWEEN
	Struct	
Unsub Event	Integer	!=,=,>=,<=,<,>
	String	LIKE,!=,=
	DateTime	>=,<=,<,>,=,BETWEEN
	Boolean	!=,=
	Struct	

Partitioning queries

In Salesforce Marketing Cloud, the Integer and DateTime datatype fields support field-based partitioning.

You can provide the additional Spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, and `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition the query.
- `LOWER_BOUND`: an **inclusive** lower bound value of the chosen partition field.

  For the timestamp field, we accept the Spark timestamp format used in Spark SQL queries.

  Examples of valid value:

  ```
  “2024-05-07T02:03:00.00Z"
  ```

- `UPPER_BOUND`: an **exclusive** upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: the number of partitions.

Example:

```
salesforcemarketingcloud_read = glueContext.create_dynamic_frame.from_options(
    connection_type="salesforcemarketingcloud",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "ListSubscriber",
        "API_VERSION": "v1",
        "PARTITION_FIELD": "CreatedDate"
        "LOWER_BOUND": "2023-09-07T02:03:00.000Z"
        "UPPER_BOUND": "2024-05-07T02:03:00.000Z"
        "NUM_PARTITIONS": "10"
    }
```

Salesforce Marketing Cloud connection options

The following are connection options for Salesforce Marketing Cloud:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your object in Salesforce Marketing Cloud.
- `API_VERSION`(String) - (Required) Used for Read. Salesforce Marketing Cloud Rest and SOAP API version you want to use.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.

- `WRITE_OPERATION`(String) - Default: INSERT. Used for write. Value should be INSERT, UPDATE, UPSERT.
- `ID_FIELD_NAMES`(String) - Default: null. Required for UPDATE/UPSERT.

Limitations and notes for Salesforce Marketing Cloud connector

The following are limitations or notes for the Salesforce Marketing Cloud connector:

- When using filter on DateTime datatype field, you need to pass the value in the format "yyyy-mm-ddThh:MM:ssZ".
- In Data Preview, the Boolean Datatype value comes as a Blank.
- For SOAP entities, you can define a maximum of two filters, and for REST entities, you can define only one filter, which restricts testing partitioning with filters.
- Several unexpected behaviors have been observed from the SaaS side: the `Link.Alias` field in the \`linksend\` entity does not support the CONTAINS operator (for example, `Link.Alias CONTAINS "ViewPrivacyPolicy"`), and filter operators for Data Extension entities (such as EQUALS and GREATER THAN) do not return the expected results.
- The SFMC ClickEvent SOAP API has a delay in reflecting newly created records so, the records created recently may not be immediately available in the API response.

  Example: If you create 5 new ClickEvent records at **2025-01-10T14:30:00** and immediately fetch them using the SOAP API, the response might not include all 5 records. It may take up to 5 minutes for the newly created records to appear in the API response. This delay can affect both data retrieval and scheduled runs as well.
- Two different DateTime formats: **2025-03-11T04:46:00** (without milliseconds) and **2025-03-11T04:46:00.000Z** are supported when performing write operations in AWS Glue (with milliseconds).
- For the Event Notification Subscription entity, a subscription can only be created for a verified callback URL, and you can have up to 200 subscriptions per callback.
- For the Event Notification Callback entity, a maximum of 50 records can be created per account.

Connecting to Salesforce Commerce Cloud

The B2C Commerce API is a collection of RESTful APIs for interacting with B2C Commerce instances. It goes by a few different names: Salesforce Commerce API, the acronym SCAPI, or just Commerce API.

The API enables developers to build a wide range of applications: from full storefronts to custom merchant tools to augment Business Manager. For all B2C Commerce customers, the API is available at no extra cost.

The API is broken into two main groups of APIs: Shopper APIs and Admin APIs. And within each group, they're divided into API families and into smaller groups focused on related functionality.

Topics

- AWS Glue support for Salesforce Commerce Cloud
- Policies containing the API operations for creating and using connections
- Configuring Salesforce Commerce Cloud
- Configuring Salesforce Commerce Cloud connections
- Reading from Salesforce Commerce Cloud entities
- Salesforce Commerce Cloud connection option reference
- Limitations

AWS Glue support for Salesforce Commerce Cloud

AWS Glue supports Salesforce Commerce Cloud as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Salesforce Commerce Cloud.

Supported as a target?

No.

Supported Salesforce Commerce Cloud API versions

v1.

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

You can also use the following managed IAM policies to allow access:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Salesforce Commerce Cloud

Before you can use AWS Glue to transfer data from Salesforce Commerce Cloud, you must meet these requirements:

Minimum requirements

- You have a Salesforce Commerce Cloud client application with clientId and clientSecret.
- Your Salesforce Commerce Cloud account is enabled for API access.

If you meet these requirements, you’re ready to connect AWS Glue to your Salesforce Commerce Cloud account. For typical connections, you don't need do anything else in Salesforce Commerce Cloud.

Configuring Salesforce Commerce Cloud connections

Salesforce Commerce Cloud supports CLIENT CREDENTIALS grant type for OAuth2.

- This grant type is considered 2-legged OAuth 2.0 as it is used by clients to obtain an access token outside of the context of a user. AWS Glue is able to use the client Id and client secret to authenticate Salesforce Commerce Cloud APIs which are provided by custom services that you define.
- Each custom service is owned by an API-Only user which has a set of roles and permissions which authorize the service to perform specific actions. An access token is associated with a single custom service.
- This grant type results in an access token which is short lived, and may be renewed by calling identity endpoint.
- For more information on Salesforce Commerce Cloud documentation on generating the Client credentials, see Salesforce documentation.

To configure an Salesforce Commerce Cloud connection:

1. In AWS Secrets Manager, create a secret with the following details. It is required to create a secret for each connection in AWS Glue.
   a. For customer managed connected app - Secret should contain the connected app Consumer Secret with USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET as key.
2. In AWS Glue Studio, create a connection under **Data Connections** by following the steps below:

   a. Under Data Connections, choose **Create connection**.

   b. When selecting a **Data Source**, select Salesforce Commerce Cloud.

   c. Provide your Salesforce Commerce Cloud **Short Code**, **Organization ID**, and **Site ID**.

   d. Select the Salesforce Commerce Cloud Domain URL of your Salesforce Commerce Cloud account.

   e. Select the IAM role which AWS Glue can assume and has permissions for following actions:

      JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

   f. Provide the OAuth scopes - optional, User Managed Client Application ClientId of the Salesforce Commerce Cloud you want to connect to.

   g. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.

   h. Select the network options if you want to use your network.

3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

4. In your AWS Glue job configuration, provide `connectionName` as an **Additional network connection**.

Reading from Salesforce Commerce Cloud entities

Prerequisites

- A Salesforce Commerce Cloud Object you would like to read from. Refer the supported entities table below to check the available entities.

Supported entities

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Assignments	Yes	Yes	Yes	Yes	Yes
Campaigns	Yes	Yes	Yes	Yes	Yes
Catalogs	Yes	Yes	Yes	Yes	Yes
Categories	Yes	Yes	Yes	Yes	Yes
Coupons	Yes	Yes	Yes	Yes	Yes
Gift Certificates	Yes	Yes	Yes	Yes	Yes
Products	Yes	Yes	Yes	Yes	Yes
Promotions	Yes	Yes	Yes	Yes	Yes
Source Code Groups	Yes	Yes	Yes	Yes	Yes

Example

```
salesforce_commerce_cloud_read = glueContext.create_dynamic_frame.from_options(
     connection_type="SalesforceCommerceCloud",
     connection_options={
         "connectionName": "connectionName",
         "ENTITY_NAME": "campaign",
         "API_VERSION": "v1"
```

```
    }
```

Salesforce Commerce Cloud entity and field details

Entities list:

- Assignments: https://developer.salesforce.com/docs/commerce/commerce-api/references/assignments
- Campaigns: https://developer.salesforce.com/docs/commerce/commerce-api/references/campaigns
- Catalogs: https://developer.salesforce.com/docs/commerce/commerce-api/references/catalogs
- Categories: https://developer.salesforce.com/docs/commerce/commerce-api/references/catalogs?meta=searchCategories
- Gift Certificates: https://developer.salesforce.com/docs/commerce/commerce-api/references/gift-certificates
- Products: https://developer.salesforce.com/docs/commerce/commerce-api/references/products
- Promotions: https://developer.salesforce.com/docs/commerce/commerce-api/references/promotions
- Source Code Groups: https://developer.salesforce.com/docs/commerce/commerce-api/references/source-code-groups

Partitioning queries

Additional spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, `NUM_PARTITIONS` can be provided if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition query.
- `LOWER_BOUND`: an inclusive lower bound value of the chosen partition field.

  For date, we accept the Spark date format used in Spark SQL queries. Example of valid values: `"2024-02-06"`.
- `UPPER_BOUND`: an exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: number of partitions.

Entity-wise partitioning field support details are captured in below table:

Entity	Partitioning Field	DataType
Campaigns	lastModified	DateTime
Campaigns	startDate	DateTime
Campaigns	endDate	DateTime
Catalogs	creationDate	DateTime
Categories	creatiionDate	DateTime
Gift Certificates	merchantId	String
Gift Certificates	creatiionDate	DateTime
Products	creatiionDate	DateTime
Products	lastModified	DateTime
Source Code Groups	creationDate	DateTime
Source Code Groups	startTime	DateTime
Source Code Groups	endTime	DateTime

Example

```
salesforceCommerceCloud_read = glueContext.create_dynamic_frame.from_options(
    connection_type="SalesforceCommerceCloud",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "coupons",
        "API_VERSION": "v1",
        "PARTITION_FIELD": "creationDate"
        "LOWER_BOUND": "2020-05-01T20:55:02.000Z"
        "UPPER_BOUND": "2024-07-11T20:55:02.000Z"
        "NUM_PARTITIONS": "10"
    }
```

Salesforce Commerce Cloud connection option reference

The following are connection options for Salesforce Commerce Cloud:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your Object in Salesforce Commerce Cloud.
- `API_VERSION`(String) - (Required) Used for Read/Write. Salesforce Commerce Cloud Rest API version you want to use. Example: v1.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.

Limitations

The following are limitations for the Salesforce Commerce Cloud connector:

- The Contains filter is not working as expected when partitioning.
- CDN Zones' entity doesn't support sandbox instances, and it supports only Development and production instance types. For more information, see https://help.salesforce.com/s/articleView?id=cc.b2c_embedded_cdn_overview.htm .
- In Salesforce Commerce Cloud, there is no API endpoint to fetch Dynamic Metadata. As a result, there is no provision to support the custom fields in the Product and Category entity.
- Site id is a mandatory query parameter. You must pass the Site Id value through the Custom Connector Setting. For more information, see Base URL and Request Formation .

- You can apply filters on maximum two fields (excluding Levels if present) in single API request with the combination of different operators as mentioned in the below table:

Filter criteria	Is supported?
One field with CONTAINS operator in single API request.	Yes
One field with Equals operator in single API request.	Yes
One field with BETWEEN operator in single API request.	Yes
Two or more fields with CONTAINS operator in single API request.	No
Two or more fields with Equals operator in single API request.	No
Two or more fields with BETWEEN operator in single API request.	No
One field with Equals and one field with CONTAINS operator in single API request.	Yes
One field with BETWEEN and one field with CONTAINS operator in single API request.	Yes
One field with BETWEEN and one field with Equals operator in single API request.	Yes
One field with Equals, one field with CONTAINS and one field with BETWEEN operator in single API request.	No
One field with Equals operator when INCREMENTAL PULL is applied in single API request.	Yes

Filter criteria	Is supported?
One field with CONTAINS operator when INCREMENTAL PULL is applied in single API request.	Yes
One field with BETWEEN operator when INCREMENTAL PULL is applied in single API request.	No
One Equals and one CONTAINS operator when INCREMENTAL PULL is applied in single API request.	No

- In some of the entities, the data type for the fields while retrieving is different from when it is used as searchable fields. As a result, there is no provision of filter feature for these fields. The following table provides the details about such fields.

Sr. No.	Entity Name	Searchable Field Name	Data Type as Searchable	Data Type as Retrievable
1	Catalog	name	String	Struct
2	Catalog	description	String	Struct
3	Category	name	String	Struct
4	Category	description	String	Struct
5	Product	name	String	Struct
6	Product	searchable	Boolean	Struct
7	Product	validFrom	DateTime	Struct
8	Product	validTo	DateTime	Struct
9	Product	type	String	Struct
10	Product	onlineFlag	Boolean	Struct

Sr. No.	Entity Name	Searchable Field Name	Data Type as Searchable	Data Type as Retrievable
11	Promotion	name	String	Struct

Connecting to Salesforce Marketing Cloud Account Engagement

Salesforce Marketing Cloud Account Engagement is a marketing automation solution that helps companies create meaningful connections, generate more pipeline, and empower sales to close more deals. If you are a Salesforce Marketing Cloud Account Engagement user, you can connect AWS Glue to your Salesforce Marketing Cloud Account Engagement account. You can use Salesforce Marketing Cloud Account Engagement as a data source in your ETL jobs. Run these jobs to transfer data from Salesforce Marketing Cloud Account Engagement to AWS services or other supported applications.

Topics

- AWS Glue support for Salesforce Marketing Cloud Account Engagement
- Policies containing the API operations for creating and using connections
- Configuring Salesforce Marketing Cloud Account Engagement
- Configuring Salesforce Marketing Cloud Account Engagement connections
- Reading from Salesforce Marketing Cloud Account Engagement entities
- Salesforce Marketing Cloud Account Engagement connection options
- Limitations and notes for Salesforce Marketing Cloud Account Engagement connector

AWS Glue support for Salesforce Marketing Cloud Account Engagement

AWS Glue supports Salesforce Marketing Cloud Account Engagement as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Salesforce Marketing Cloud Account Engagement in either Async or Sync mode.

Supported as a target?

No.

Supported Salesforce Marketing Cloud Account Engagement API versions

The following Salesforce Marketing Cloud Account Engagement API versions are supported:

- v5

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention

for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Salesforce Marketing Cloud Account Engagement

Before you can use AWS Glue to transfer data from Salesforce Marketing Cloud Account Engagement, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- You have a Salesforce marketing account.
- You have a licensed Account Engagement plan for the Salesforce account.
- You have synced the Salesforce user with the Account Engagement user.
- You have created a new connected app under App Manager to obtain OAuth Credentials.

If you meet these requirements, you're ready to connect AWS Glue to your Salesforce Marketing Cloud Account Engagement account.

Configuring Salesforce Marketing Cloud Account Engagement connections

The grant type determines how AWS Glue communicates with Salesforce Marketing Cloud Account Engagement to request access to your data. Your choice affects the requirements that you must meet before you create the connection. Salesforce Marketing Cloud Account Engagement supports only the AUTHORIZATION_CODE grant type for OAuth 2.0.

- This grant type is considered "three-legged" OAuth as it relies on redirecting users to a third-party authorization server to authenticate the user. It is used when creating connections via the AWS Glue console.
- Users may still opt to create their own connected app in Salesforce Marketing Cloud Account Engagement and provide their own client ID and client secret when creating connections through the AWS Glue console. In this scenario, they will still be redirected to Salesforce Marketing Cloud Account Engagement to login and authorize AWS Glue to access their resources.
- This grant type results in a refresh token and access token. The access token is short lived, and may be refreshed automatically without user interaction using the refresh token.

- For public Salesforce Marketing Cloud Account Engagement documentation on creating a connected app for Authorization Code OAuth flow, see Authentication.

To configure a Salesforce Marketing Cloud Account Engagement connection:

1. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
    a. When selecting a **Connection type**, select Salesforce Marketing Cloud Account Engagement.
    b. Provide the `INSTANCE_URL` of the Salesforce Marketing Cloud Account Engagement instance you want to connect to.
    c. Provide the `PARDOT_BUSINESS_UNIT_ID` of the Salesforce Marketing Cloud Account Engagement instance you want to connect to.
    d. Select the appropriate **Authorization Code URL** from the dropdown.
    e. Select the appropriate **Token URL** from the dropdown.
    f. Select the AWS IAM role which AWS Glue can assume and has permissions for following actions:

       JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

    g. Provide the User Managed Client Application Client ID (the client ID from the connected app).

    h. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens. The selected secret needs to have a key `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` with the value being the Client Secret from the connected app.
    i. Select the network options if you want to use your network.
2. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.
3. In your AWS Glue job configuration, provide `connectionName` as an **Additional network connection**.

Reading from Salesforce Marketing Cloud Account Engagement entities

Prerequisite

A Salesforce Marketing Cloud Account Engagement object you would like to read from. You will need the object name.

Supported entities for Sync source:

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Campaign	Yes	Yes	Yes	Yes	Yes
Dynamic Content	Yes	Yes	Yes	Yes	Yes
Email	Yes	Yes	Yes	Yes	Yes
Email Template	Yes	Yes	Yes	Yes	Yes
Engagement Studio Program	Yes	Yes	Yes	Yes	Yes
Folder Contents	Yes	Yes	Yes	Yes	Yes

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Landing Page	Yes	Yes	Yes	Yes	Yes
Lifecycle History	Yes	Yes	Yes	Yes	Yes
Lifecycle Stage	Yes	Yes	Yes	Yes	Yes
List	Yes	Yes	Yes	Yes	Yes
List Email	Yes	Yes	Yes	Yes	Yes
List Membership	Yes	Yes	Yes	Yes	Yes
Opportunity	Yes	Yes	Yes	Yes	Yes
Prospect	Yes	Yes	Yes	Yes	Yes
Prospect Account	Yes	Yes	Yes	Yes	Yes
User	Yes	Yes	Yes	Yes	Yes

Example:

```
salesforcepardot_read = glueContext.create_dynamic_frame.from_options(
    connection_type="SalesforcePardot",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "entityName",
        "API_VERSION": "v5"
    }
   )
```

Supported entities for Async source:

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Campaign	Yes	No	No	Yes	No
Dynamic Content	Yes	No	No	Yes	No
Email Template	Yes	No	No	Yes	No
Landing Page	Yes	No	No	Yes	No
Lifecycle History	Yes	No	No	Yes	No
Lifecycle Stage	Yes	No	No	Yes	No
List	Yes	No	No	Yes	No
List Email	Yes	No	No	Yes	No
List Membership	Yes	No	No	Yes	No
Opportunity	Yes	No	No	Yes	No
Prospect	Yes	No	No	Yes	No
Prospect Account	Yes	No	No	Yes	No
User	Yes	No	No	Yes	No

Example:

```
salesforcepardot_read = glueContext.create_dynamic_frame.from_options(
    connection_type="SalesforcePardot",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "entityName",
        "API_VERSION": "v5",
        "TRANSFER_MODE": "ASYNC"
    }
   )
```

Salesforce Marketing Cloud Account Engagement entity and field details:

To view the field details for thhe following entities, navigate to Salesforce Marketing Cloud Account Engagement API, choose **Guides**, scroll down to **Open Source API Wrappers**, expand **Version 5 Docs** from the menu and choose an entity.

Entities list:

- Campaign
- Dynamic Content
- Email
- Email Template
- Engagement Studio Program
- Folder Content
- Landing Page
- Lifecycle History
- Lifecycle Stage
- List
- List Email
- List Membership
- Opportunity
- Prospect
- Prospect Account
- User

In addition to the fields mentioned above, the Async mode supports specific filterable fields for each entity, as shown in the table below.

Entity	Additional filterable fields supported in Async
Campaign	`createdAfter` , `createdBefore` , `deleted`, `updatedAfter` , `updatedBefore`
Dynamic Content	`createdAfter` , `createdBefore` , `deleted`, `updatedAfter` , `updatedBefore`
Email Template	`createdAfter` , `createdBefore` , `deleted`, `updatedAfter` , `updatedBefore`
Engagement Studio Program	-
Landing Page	`createdAfter` , `createdBefore` , `deleted`, `updatedAfter` , `updatedBefore`
Lifecycle History	`createdAfter` , `createdBefore`
Lifecycle Stage	`createdAfter` , `createdBefore` , `deleted`, `updatedAfter` , `updatedBefore`
List	`createdAfter` , `createdBefore` , `deleted`, `updatedAfter` , `updatedBefore`
List Email	`createdAfter` , `createdBefore` , `deleted`, `updatedAfter` , `updatedBefore`
List Membership	`createdAfter` , `createdBefore` , `deleted`, `updatedAfter` , `updatedBefore`

Entity	Additional filterable fields supported in Async
Opportunity	`createdAfter` , `createdBefore` , `deleted`, `updatedAfter` , `updatedBefore`
Prospect	`createdAfter` , `createdBefore` , `deleted`, `updatedAfter` , `updatedBefore`
Prospect Account	`createdAfter` , `createdBefore` , `deleted`
User	`createdAfter` , `createdBefore` , `deleted`, `updatedAfter` , `updatedBefore`

For more information about the additional fields, refer to Salesforce Export API

Note the following considerations for the connector:

- The value of the `delete` field in the entities can be `false` (default), `true`, or `all`.

Partitioning queries

Filter-based partitioning:

You can provide the additional Spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, and `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition the query.
- `LOWER_BOUND`: an **inclusive** lower bound value of the chosen partition field.

  For the Datetime field, we accept the Spark timestamp format used in SPark SQL queries.

  Examples of valid value:

```
"2022-01-01T01:01:01.000Z"
```

- `UPPER_BOUND`: an **exclusive** upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: the number of partitions.
- `PARTITION_BY`: the type of partitioning to be performed. "FIELD" is to be passed in case of field-based partitioning.

Example:

```
salesforcepardot_read = glueContext.create_dynamic_frame.from_options(
    connection_type="salesforcepardot",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "entityName",
        "API_VERSION": "v5",
        "PARTITION_FIELD": "createdAt"
        "LOWER_BOUND": "2022-01-01T01:01:01.000Z"
        "UPPER_BOUND": "2024-01-01T01:01:01.000Z"
        "NUM_PARTITIONS": "10",
        "PARTITION_BY": "FIELD"
    }
   )
```

Salesforce Marketing Cloud Account Engagement connection options

The following are connection options for Salesforce Marketing Cloud Account Engagement:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your object in Salesforce Marketing Cloud Account Engagement.
- `PARDOT_BUSINESS_UNIT_ID` - (Required) Used for creating a connection. The business unit ID of the Salesforce Marketing Cloud Account Engagement instance you want to connect to.
- `API_VERSION`(String) - (Required) Used for Read. Salesforce Marketing Cloud Account Engagement Rest API version you want to use.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) -
  - In Sync mode - Default: empty. Used for Read. It should be in the Spark SQL format.

  - In Async mode - Default : Current `DateTime` value (as per user's timezone) - 1 year. Used for Read.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.
- `INSTANCE_URL`(String) - (Required) Used for Read. A valid Salesforce Marketing Cloud Account Engagement instance URL.
- `PARTITION_BY`(String) - (Required) Used for Read. The type of partitioning to be performed. "FIELD" is to be passed in case of field-based partitioning.
- `TRANSFER_MODE`(String) - (Optional), Value to be used for running a job in ASYNC mode , if this option not provided job will run in SYNC mode.

Limitations and notes for Salesforce Marketing Cloud Account Engagement connector

The following notes and limitations apply:

- When both a limit and partitioning are applied, the limit takes precedence over the partitioning.
- As per the API Docs, `SalesforceMarketingCloudEngagement` enforced a RateLimit on daily and concurrent requests. For more information, see Rate Limits.
- The Export API is subject to the daily Account Engagement API call limit and the concurrent Account Engagement API call limit for your account.
- Similar to a queue, Export/Async API calls are executed sequentially for each account. Older exports are processed before newer exports.
- Partition is not supported in Async mode.
- The number of selected fields specified in the Export/Async API calls can't exceed 150.
- The **Prospect** entity supports over 150 fields, but only 150 fields can be selected at a time. If `Select All` is chosen, some fields will be excluded. To retrieve data for these excluded fields, you have to include them in the `Selected Fields` option.

The following is the list of excluded fields in `SELECT_ALL` - `updatedBy.firstName`, `updatedBy.lastName`, `updatedBy.jobTitle`, `updatedBy.roleName`, `updatedBy.salesforceId`, `updatedBy.createdAt`, `updatedBy.updatedAt`, `updatedBy.isDeleted`, `updatedBy.createdById`, `updatedBy.updatedById`, `updatedBy.tagReplacementLanguage`

- Collection fields can't be exported for Async. For example, on List Email the `senderOptions` and `replyToOptions` fields are not supported.
- For all entities, filter is mandatory. If no filter is provided the default filter predicate is set to the `Created After` field with a value of the current date-time (adjusted to your time zone) minus one year.
- As per Salesforce Marketing Cloud Account Engagement limitations, in Async, the maximum range to fetch data is 1 year. If a query is provided for more than 1 year, the job will throw an error.
- Currently, there is a bug in Salesforce Pardot. When the job includes only a single field which does not have any data, the field value is not returning correct result and instead, the field name is being returned multiple times. The Salesforce Pardot team is aware of the issue and is actively working on a resolution.

Connecting to SAP HANA in AWS Glue Studio

AWS Glue provides built-in support for SAP HANA. AWS Glue Studio provides a visual interface to connect to SAP HANA, author data integration jobs, and run them on the AWS Glue Studio serverless Spark runtime.

AWS Glue Studio creates a unified connection for SAP HANA. For more information, see Considerations.

Topics

- Creating a SAP HANA connection
- Creating a SAP HANA source node
- Creating a SAP HANA target node
- Advanced options

Creating a SAP HANA connection

To connect to SAP HANA from AWS Glue, you will need to create and store your SAP HANA credentials in a AWS Secrets Manager secret, then associate that secret with a SAP HANA AWS Glue connection. You will need to configure network connectivity between your SAP HANA service and AWS Glue.

Prerequisites:

- If your SAP HANA service is in an Amazon VPC, configure Amazon VPC to allow your AWS Glue job to communicate with the SAP HANA service without traffic traversing the public internet.

  In Amazon VPC, identify or create a **VPC**, **Subnet** and **Security group** that AWS Glue will use while executing the job. Additionally, you need to ensure Amazon VPC is configured to permit network traffic between your SAP HANA endpoint and this location. Your job will need to establish a TCP connection with your SAP HANA JDBC port. For more information about SAP HANA ports, see the SAP HANA documentation. Based on your network layout, this may require changes to security group rules, Network ACLs, NAT Gateways and Peering connections.

To configure a connection to SAP HANA:

1. In AWS Secrets Manager, create a secret using your SAP HANA credentials. To create a secret in Secrets Manager, follow the tutorial available in Create an AWS Secrets Manager secret in the AWS Secrets Manager documentation. After creating the secret, keep the Secret name, *`secretName`* for the next step.

   - When selecting **Key/value pairs**, create a pair for the key `user` with the value *`saphanaUsername`*.
   - When selecting **Key/value pairs**, create a pair for the key `password` with the value *`saphanaPassword`*.

2. In the AWS Glue console, create a connection by following the steps in the section called “Adding an AWS Glue connection”. After creating the connection, keep the connection name, *`connectionName`*, for future use in AWS Glue.

   - When selecting a **Connection type**, select SAP HANA.
   - When providing **SAP HANA URL**, provide the URL for your instance.

SAP HANA JDBC URLs are in the form `jdbc:sap://saphanaHostname:saphanaPort/?databaseName=saphanaDBname,Paramete`

AWS Glue requires the following JDBC URL parameters:

- `databaseName` – A default database in SAP HANA to connect to.

- When selecting an **AWS Secret**, provide *`secretName`*.

After creating a AWS Glue SAP HANA connection, you will need to perform the following steps before running your AWS Glue job:

- Grant the IAM role associated with your AWS Glue job permission to read *`secretName`*.

Creating a SAP HANA source node

Prerequisites needed

- An AWS Glue SAP HANA connection, configured with an AWS Secrets Manager secret, as described in the previous section, the section called “Creating a SAP HANA connection”.
- Appropriate permissions on your job to read the secret used by the connection.
- A SAP HANA table you would like to read from, *`tableName`*, or query *`targetQuery`*.

  A table can be specified with a SAP HANA table name and schema name, in the form *`schemaName`*`.`*`tableName`*. The schema name and "." separator are not required if the table is in the default schema, "public". Call this *`tableIdentifier`*. Note that the database is provided as a JDBC URL parameter in `connectionName`.

Adding a SAP HANA data source

To add a Data source – SAP HANA node:

1. Choose the connection for your SAP HANA data source. Since you have created it, it should be available in the dropdown. If you need to create a connection, choose **Create SAP HANA connection**. For more information see the previous section, the section called “Creating a SAP HANA connection”.

   Once you have chosen a connection, you can view the connection properties by clicking **View properties**.

2. Choose a **SAP HANA Source** option:

    - **Choose a single table** – access all data from a single table.
    - **Enter custom query** – access a dataset from multiple tables based on your custom query.

3. If you chose a single table, enter *`tableName`*.

    If you chose **Enter custom query**, enter a SQL SELECT query.

4. In **Custom SAP HANA properties**, enter parameters and values as needed.

Creating a SAP HANA target node

Prerequisites needed

- A AWS Glue SAP HANA connection, configured with an AWS Secrets Manager secret, as described in the previous section, the section called "Creating a SAP HANA connection".
- Appropriate permissions on your job to read the secret used by the connection.
- A SAP HANA table you would like to write to, *`tableName`*.

  A table can be specified with a SAP HANA table name and schema name, in the form *`schemaName`*`.`*`tableName`*. The schema name and "." separator are not required if the table is in the default schema, "public". Call this *`tableIdentifier`*. Note that the database is provided as a JDBC URL parameter in `connectionName`.

Adding a SAP HANA data target

To add a Data target – SAP HANA node:

1. Choose the connection for your SAP HANA data source. Since you have created it, it should be available in the dropdown. If you need to create a connection, choose **Create SAP HANA connection**. For more information see the previous section, the section called "Creating a SAP HANA connection".

    Once you have chosen a connection, you can view the connection properties by clicking **View properties**.

2. Configure **Table name** by providing *`tableName`*.
3. In **Custom Teradata properties**, enter parameters and values as needed.

Advanced options

You can provide advanced options when creating a SAP HANA node. These options are the same as those available when programming AWS Glue for Spark scripts.

See the section called "SAP HANA connections".

Connecting to SAP OData

SAP OData is a standard Web protocol used for querying and updating data present in SAP using ABAP (Advanced Business Application Programming), applying and building on Web technologies such as HTTP to provide access to information from a variety of external applications, platforms and devices. With the product, you can access everything you need to help you seamlessly integrate with your SAP system, application, or data.

Topics

- AWS Glue support for SAP OData
- Create connections
- Creating SAP OData job
- Using the SAP OData state management script
- Partitioning for Non ODP entities
- SAP OData connection options
- SAP OData entity and field details

AWS Glue support for SAP OData

AWS Glue supports SAP OData as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from SAP OData.

Supported SAP OData API versions

The following SAP OData API versions are supported:

- 2.0

Supported sources

The following sources are supported:

- ODP (Operational Data Provisioning) Sources:
  - BW Extractors (DataSources)
  - CDS Views
  - SLT
- Non-ODP Sources, for example:
  - CDS View Services
  - RFC-based Services
  - Custom ABAP Services

Supported SAP Components

The following are minimum requirements:

- You must enable catalog service for service discovery.
  - Configure operational data provisioning (ODP) data sources for extraction in the SAP Gateway of your SAP system.
  - **OData V2.0**: Enable the OData V2.0 catalog service(s) in your SAP Gateway via transaction `/IWFND/MAINT_SERVICE`.
  - Enable OData V2.0 services in your SAP Gateway via transaction `/IWFND/MAINT_SERVICE`.
  - Your SAP OData service must support client side pagination/query options such as `$top` and `$skip`. It must also support system query option `$count`.
  - You must provide the required authorization for the user in SAP to discover the services and extract data using SAP OData services. Refer to the security documentation provided by SAP.
- If you want to use OAuth 2.0 as an authorization mechanism, you must enable OAuth 2.0 for the OData service and register the OAuth client per SAP documentation.
- To generate an OData service based on ODP data sources, SAP Gateway Foundation must be installed locally in your ERP/BW stack or in a hub configuration.
  - For your ERP/BW applications, the SAP NetWeaver AS ABAP stack must be at 7.50 SP02 or above.
  - For the hub system (SAP Gateway), the SAP NetWeaver AS ABAP of the hub system must be 7.50 SP01 or above for remote hub setup.
- For non-ODP sources, your SAP NetWeaver stack version must be 7.40 SP02 or above.

Supported Authentication Methods

The following authentication methods are supported:

- Basic Authentication
- OAuth 2.0

Prerequisites

Prior to initiating an AWS Glue job for data extraction from SAP OData using the SAP OData connection, complete the following prerequisites:

- The relevant SAP OData Service must be activated in the SAP system, ensuring the data source is available for consumption. If the OData service is not activated, the Glue job will not be able to access or extract data from SAP.
- Appropriate authentication mechanisms such as basic (custom) authentication or OAuth 2.0 must be configured in SAP to ensure that the AWS Glue job can successfully establish a connection with the SAP OData service.
- Configure IAM policies to grant the AWS Glue job appropriate permissions for accessing SAP, Secrets Manager, and other AWS resources involved in the process.
- If the SAP system is hosted within a private network, VPC connectivity must be configured to ensure that the AWS Glue job can securely communicate with SAP without exposing sensitive data over public internet.

AWS Secrets Manager can be used to securely store sensitive information such as SAP credentials, which the AWS Glue job can dynamically retrieve at runtime. This approach eliminates the need to hard-code credentials, enhancing security and flexibility.

The following prerequisites provide step-by-step guidance on how to set up each component for a smooth integration between AWS Glue and SAP OData.

Topics

- SAP OData activation
- IAM policies
- Connectivity / VPC Connection
- SAP Authentication

- AWS Secrets Manager to store your Auth secret

SAP OData activation

Complete the following steps for SAP OData connection:

ODP Sources

Before you can transfer data from an ODP provider, you must meet the following requirements:

- You have an SAP NetWeaver AS ABAP instance.
- Your SAP NetWeaver instance contains an ODP provider that you want to transfer data from. ODP providers include:
  - SAP DataSources (Transaction code RSO2)
  - SAP Core Data Services ABAP CDS Views
  - SAP BW or SAP BW/4HANA systems (InfoObject, DataStore Object)
  - Real-time replication of Tables and DB-Views from SAP Source System via SAP Landscape Replication Server (SAP SLT)
  - SAP HANA Information Views in SAP ABAP based Sources
- Your SAP NetWeaver instance has the SAP Gateway Foundation component.
- You have created an OData service that extracts data from your ODP provider. To create the OData service, you use the SAP Gateway Service Builder. To access your ODP data, Amazon AppFlow calls this service by using the OData API. For more information, see Generating a Service for Extracting ODP Data via OData in the SAP BW/4HANA documentation.
- To generate an OData service based on ODP data sources, SAP Gateway Foundation must be installed locally in your ERP/BW stack or in a hub configuration.
  - For your ERP/BW applications, the SAP NetWeaver AS ABAP stack must be at 7.50 SP02 or above.
  - For the hub system (SAP Gateway), the SAP NetWeaver AS ABAP of the hub system must be 7.50 SP01 or above for remote hub setup.

Non-ODP Sources

- Your SAP NetWeaver stack version must be 7.40 SP02 or above.
- You must enable catalog service for service discovery.

  - **OData V2.0**: The OData V2.0 catalog service(s) can be enabled in your SAP Gateway via transaction `/IWFND/MAINT_SERVICE`
- Your SAP OData service must support client side pagination/query options such as `$top` and `$skip`. It must also support system query option `$count`.
- For OAuth 2.0, you must enable OAuth 2.0 for the OData service and register the OAuth client per SAP documentation and set the authorized redirect URL as follows:
  - `https://<region>.console.aws.amazon.com/gluestudio/oauth`, replacing `<region>` with the region where AWS Glue is running, example: us-east-1.
  - You must enable secure setup for connecting over HTTPS.
- You must provide required authorization for the user in SAP to discover the services and extract data using SAP OData services. Please refer to the security documentation provided by SAP.

IAM policies

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    },
    {
      "Effect": "Allow",
      "Action": [
          "secretsmanager:DescribeSecret",
          "secretsmanager:GetSecretValue",
          "secretsmanager:PutSecretValue"
        ],
        "Resource": "*"
      }
    ]
  }
```

The role must grant access to all the resources used by the job, for example Amazon S3. If you don't want to use the above method, alternatively use the following managed IAM policies.

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.
- SecretsManagerReadWrite – Provides read/write access to AWS Secrets Manager via the AWS Management Console. Note: this excludes IAM actions, so combine with `IAMFullAccess` if rotation configuration is required.

IAM Policies/Permissions needed to configure VPC

The following IAM permissions are required while using VPC connection for creating AWS Glue Connection. For more details, refer to create an IAM policy for AWS Glue.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
          "ec2:CreateNetworkInterface",
          "ec2:DeleteNetworkInterface",
          "ec2:DescribeNetworkInterfaces"
        ],
        "Resource": [
          "*"
        ]
      }
    ]
  }
```

Connectivity / VPC Connection

Steps for VPC Connection:

1. Use existing VPC connection or create a new connection by following the Amazon VPC documentation.
2. Make sure you have NAT Gateway which routes the traffic to internet.
3. Choose VPC endpoint as Amazon S3 Gateway to create connection.
4. Enable DNS resolution and DNS hostname to use AWS provided DNS Services.
5. Go to created VPC and add necessary endpoints for different services like STS, AWS Glue, Secret Managers.
    a. Choose Create Endpoint.
    b. For Service Category, choose AWS Services.
    c. For Service Name, choose the service that you are connecting to.
    d. Choose VPC and Enable DNS Name.
    e. VCP Endpoints required for VPC connection:
        i. STS
        ii. AWS Glue
        iii. Secrets Manager

Security Group Configuration

Security group must allow traffic to its listening port from AWS Glue VPC for AWS Glue to be able to connect to it. It is a good practice to restrict the range of source IP addresses as much as possible.

AWS Glue requires special security group that allows all inbound traffic from itself. You can create a self-referencing rule that allows all traffic originating from the security group. You can modify an existing security group and specify the security group as source.

Open the communication from the HTTPS ports of the URL endpoint (either NLB or SAP instance).

Connectivity options

- HTTPS connection with internal and external NLB, SSL certificate from certificate authority (CA), not self-signed SSL certificate
- HTTPS connection with SAP instance SSL certificate from certificate authority (CA), not self-signed SSL certificate

SAP Authentication

The SAP connector supports both CUSTOM (this is SAP BASIC authentication) and OAUTH authentication methods.

Custom Authentication

AWS Glue supports Custom (Basic Authentication) as a method for establishing connections to your SAP systems, allowing the use of a username and password for secure access. This auth type works well for automation scenarios as it allows using username and password up front with the permissions of a particular user in the SAP OData instance. AWS Glue is able to use the username and password to authenticate SAP OData APIs. In AWS Glue, basic authorization is implemented as custom authorization.

For public SAP OData documentation for Basic Auth flow, see HTTP Basic Authentication.

OAuth 2.0 Authentication

AWS Glue also supports OAuth 2.0 as a secure authentication mechanism for establishing connections to your SAP systems. This enables seamless integration while ensuring compliance with modern authentication standards and enhancing the security of data access.

AUTHORIZATION_CODE Grant Type

The grant type determines how AWS Glue communicates with SAP OData to request access to your data. SAP OData supports only the `AUTHORIZATION_CODE` grant type. This grant type is considered "three-legged" OAuth as it relies on redirecting users to the third-party authorization server to authenticate the user. It is used when creating connections via the AWS Glue console.

Users may still opt to create their own connected app in SAP OData and provide their own client ID and client secret when creating connections through the AWS Glue console. In this scenario, they will still be redirected to SAP OData to login and authorize AWS Glue to access their resources.

This grant type results in a refresh token and access token. The access token is short lived, and may be refreshed automatically without user interaction using the refresh token.

For public SAP OData documentation on creating a connected app for Authorization Code OAuth flow, see Authentication Using OAuth 2.0.

AWS Secrets Manager to store your Auth secret

You will need to store the SAP OData connection secrets in AWS Secrets Manager, configure the necessary permissions for retrieval as specified in the IAM policies section, and use it while creating a connection.

Use the AWS Management Console for AWS Secrets Manager to create a secret for your SAP source. For more information, see Create an AWS Secrets Manager secret. Details in AWS Secrets Manager should include the elements in the following code.

Custom Authentication Secret

You will need to enter your SAP system username in place of *<your SAP username>* and its password in place of *<your SAP username password>* and True or False. In this context, setting `basicAuthDisableSSO` to `true` disables Single Sign-On (SSO) for Basic Authentication requests, requiring explicit user credentials for each request. Conversely, setting it to `false` allows the use of existing SSO sessions if available.

```
{
   "basicAuthUsername": "<your SAP username>",
   "basicAuthPassword": "<your SAP username password>",
   "basicAuthDisableSSO": "<True/False>",
   "customAuthenticationType": "CustomBasicAuth"
}
```

OAuth 2.0 Secret

In case you are using OAuth 2.0 as your authentication mechanism, the secret in the AWS Secrets Manager should have the **User Managed Client Application ClientId** in the following format. You will need to enter your SAP client secret in place of <your client secret>.

```
{"USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET": "<your client secret>"
}
```

Create connections

To configure an SAP OData connection:

1. Sign in to the AWS Management Console and open the AWS Glue console. In the AWS Glue Studio, create a connection by following the steps below:
    a. Click Data connections on the left panel.
    b. Click on Create connection.
    c. Select **SAP OData** in **Choose data source**
    d. Provide the **Application host URL** of the SAP OData instance you want to connect to. This application host url must be accessible over public internet for non VPC connection.
    e. Provide the **Application service path** of the SAP OData instance you want to connect to. This is the same as the catalog service path. For example: `/sap/opu/odata/iwfnd/catalogservice;v=2`. AWS Glue doesn't accept specific object path.
    f. Provide the **Client number** of the SAP OData instance you want to connect to. Acceptable values are [001-999]. Example: 010
    g. Provide the **Port number** of the SAP OData instance you want to connect to. Example: 443
    h. Provide the **Logon language** of the SAP OData instance you want to connect to. Example: EN
    i. Select the AWS IAM role which AWS Glue can assume and has permissions as outlined in the IAM policies section.
    j. Select the **Authentication Type** which you want to use for this connection in AWS Glue from the dropdown list: OAUTH2 or CUSTOM
        i. CUSTOM - Select the secret you created as specified in the AWS Secrets Manager to store your Auth secret section.
        ii. OAUTH 2.0 - enter the following inputs only in case of OAuth 2.0:
            A. Under **User Managed Client Application ClientId**, enter your client id.
            B. `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` (your client secret) in the AWS Secrets Manager that you created in the AWS Secrets Manager to store your Auth secret section.
            C. Under **Authorization Code URL**, enter your authorization code URL.
            D. Under **Authorization Tokens URL**, enter your authorization token URL.

   E. Under **OAuth Scopes**, enter your OAuth scopes separated by space. Example: `/IWFND/SG_MED_CATALOG_0002 ZAPI_SALES_ORDER_SRV_0001`

  k. Select the network options if you want to use your network. For more details, see Connectivity / VPC Connection.

2. Grant the IAM role associated with your AWS Glue job permission to read `secretName`. For more details, see the IAM policies.
3. Choose **Test connection** and test your connection. If the connection test passes, click next, enter your connection name and save your connection. Test connection functionality is not available if you have chosen Network options (VPC).

Creating SAP OData job

Refer to Building visual ETL jobs with AWS Glue Studio

Operational Data Provisioning (ODP) Sources

Operational Data Provisioning (ODP) provides a technical infrastructure that you can use to support data extraction and replication for various target applications and supports delta mechanisms in these scenarios. In case of a delta procedure, the data from a source (ODP Provider) is automatically written to a delta queue (Operational Delta Queue – ODQ) using an update process or passed to the delta queue using an extractor interface. An ODP Provider can be a DataSource (extractors), ABAP Core Data Services Views (ABAP CDS Views), SAP BW or SAP BW/4HANA, SAP Landscape Transformation Replication Server (SLT), and SAP HANA Information Views (calculation views). The target applications (referred to as ODQ 'subscribers' or more generally “ODP Consumers”) retrieve the data from the delta queue and continue processing the data.

Full Load

In the context of SAP OData and ODP entities, a **Full Load** refers to the process of extracting all available data from an ODP entity in a single operation. This operation retrieves the complete dataset from the source system, ensuring that the target system has a comprehensive and up-to-date copy of the entity's data. Full loads are typically used for sources that do not support incremental loads or when a refresh of the target system is required.

Example

You can explicitly set the ENABLE_CDC flag to false, when creating the DynamicFrame. Note: ENABLE_CDC is false by default, if you don't want to initialize the delta queue, you don't have to send this flag or set it to true. Not setting this flag to true will result in a full load extraction.

```
sapodata_df = glueContext.create_dynamic_frame.from_options(
    connection_type="SAPOData",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "entityName",
        "ENABLE_CDC": "false"
    }, transformation_ctx=key)
```

Incremental Load

An **incremental load** in the context of ODP (Operational Data Provisioning) entities involves extracting only the new or changed data (deltas) from the source system since the last data extraction, avoiding preprocessing the already processed records. This approach significantly improves efficiency, reduces data transfer volumes, enhances performance, ensures efficient synchronization between systems, and minimizes processing time, especially for large datasets that change frequently.

Delta Token based Incremental Transfers

To enable Incremental Transfer using Change Data Capture (CDC) for ODP-enabled entities that support it, follow these steps:

1. Create the Incremental Transfer job in script mode.
2. When creating the DataFrame or Glue DynamicFrame, you need to pass the option `"ENABLE_CDC": "True"`. This option ensures that you will receive a Delta Token from SAP, which can be used for subsequent retrieval of changed data.

The delta token will be present in the last row of the dataframe, in the DELTA_TOKEN column. This token can be used as a connector option in subsequent calls to incrementally retrieve the next set of data.

Example

- We set the ENABLE_CDC flag to `true`, when creating the DynamicFrame. Note: ENABLE_CDC is `false` by default, if you don't want to initialize the delta queue, you don't need to send this flag or set it to true. Not setting this flag to true will result in a full load extraction.

```
sapodata_df = glueContext.create_dynamic_frame.from_options(
    connection_type="SAPOData",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "entityName",
        "ENABLE_CDC": "true"
    }, transformation_ctx=key)

# Extract the delta token from the last row of the DELTA_TOKEN column
delta_token_1 = your_logic_to_extract_delta_token(sapodata_df) # e.g.,
 D20241029164449_000370000
```

- The extracted delta token can be passed as a an option to retrieve new events.

```
sapodata_df_2 = glueContext.create_dynamic_frame.from_options(
    connection_type="SAPOData",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "entityName",
        // passing the delta token retrieved in the last run
        "DELTA_TOKEN": delta_token_1
    } , transformation_ctx=key)

# Extract the new delta token for the next run
delta_token_2 = your_logic_to_extract_delta_token(sapodata_df_2)
```

Note that the last record, in which the `DELTA_TOKEN` is present, is not a transactional record from source, and is only there for the purpose of passing the delta token value.

Apart from the `DELTA_TOKEN`, the following fields are returned in each row of the dataframe.

- **GLUE_FETCH_SQ**: This is a sequence field, generated from the EPOC timestamp in the order the record was received, and is unique for each record. This can be used if you need to know or establish the order of changes in the source system. This field will be present only for ODP enabled entities.
- **DML_STATUS**: This will show `UPDATED` for all newly inserted and updated records from the source, and `DELETED` for records that have been deleted from source.

For more details about how to manage state and reuse the delta token to retrieve changed records through an example refer to the Using the SAP OData state management script section.

Delta Token Invalidation

A delta token is associated with the service collection and a user. If a new initial pull with `“ENABLE_CDC” : “true”` is initiated for the same service collection and the user, all previous delta tokens issued as a result of a previous initialization will be invalidated by SAP OData service. Invoking the connector with an expired delta token will lead to an exception:

```
Could not open data access via extraction API RODPS_REPL_ODP_OPEN
```

OData Services (Non-ODP Sources)

Full Load

For Non-ODP (Operational Data Provisioning) systems, a **Full Load** involves extracting the entire dataset from the source system and loading it into the target system. Since Non-ODP systems do not inherently support advanced data extraction mechanisms like deltas, the process is straightforward but can be resource-intensive depending on the size of the data.

Incremental Load

For systems or entities that do not support **ODP (Operational Data Provisioning)**, incremental data transfer can be managed manually by implementing a timestamp based mechanism to track and extract changes.

Timestamp based Incremental Transfers

For non-ODP enabled entities(or for ODP enabled entities that don’t use the ENABLE_CDC flag), we can use a `filteringExpression` option in the connector to indicate the `datetime` interval for which we want to retrieve data. This method relies on a timestamp field in you data that represents when each record was last created/modified.

Example

Retrieving records that changed after 2024-01-01T00:00:00.000

```
sapodata_df = glueContext.create_dynamic_frame.from_options(
    connection_type="SAPOData",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "entityName",
        "filteringExpression": "LastChangeDateTime >= 2024-01-01T00:00:00.000"
    }, transformation_ctx=key)
```

Note: In this example, `LastChangeDateTime` is the field that represents when each record was last modified. The actual field name may vary depending on your specific SAP OData entity.

To get a new subset of data in subsequent runs, you would update the `filteringExpression` with a new timestamp. Typically, this would be the maximum timestamp value from the previously retrieved data.

Example

```
max_timestamp = get_max_timestamp(sapodata_df)  # Function to get the max timestamp
 from the previous run
next_filtering_expression = f"LastChangeDateTime > {max_timestamp}"

# Use this next_filtering_expression in your next run
```

In the next section, we will provide an automated approach to manage these timestamp-based incremental transfers, eliminating the need to manually update the filtering expression between runs.

Using the SAP OData state management script

To use the SAP OData state management script in your AWS Glue job, follow these steps:

- Download the state management script: `s3://aws-blogs-artifacts-public/artifacts/BDB-4789/sap_odata_state_management.zip` from the public Amazon S3 bucket.
- Upload the script to an Amazon S3 bucket that your AWS Glue job has permissions to access.
- Reference the script in your AWS Glue job: When creating or updating your AWS Glue job, pass the `'--extra-py-files'` option referencing the script path in your Amazon S3 bucket. For example: `--extra-py-files s3://your-bucket/path/to/sap_odata_state_management.py`
- Import and use the state management library in your AWS Glue job scripts.

Delta-token based Incremental Transfer example

Here's an example of how to use the state management script for delta-token based incremental transfers:

```
from sap_odata_state_management import StateManagerFactory, StateManagerType, StateType

# Initialize the state manager
state_manager = StateManagerFactory.create_manager(
    manager_type=StateManagerType.JOB_TAG,
    state_type=StateType.DELTA_TOKEN,
    options={
        "job_name": args['JOB_NAME'],
        "logger": logger
    }
)

# Get connector options (including delta token if available)
key = "SAPODataNode"
connector_options = state_manager.get_connector_options(key)

# Use the connector options in your Glue job
df = glueContext.create_dynamic_frame.from_options(
    connection_type="SAPOData",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "entityName",
        "ENABLE_CDC": "true",
        **connector_options
    }
)

# Process your data here...

# Update the state after processing
state_manager.update_state(key, sapodata_df.toDF())
```

Timestamp based Incremental Transfer example

Here's an example of how to use the state management script for delta-token based incremental transfers:

```
from sap_odata_state_management import StateManagerFactory, StateManagerType, StateType
```

```
# Initialize the state manager
state_manager = StateManagerFactory.create_manager(
    manager_type=StateManagerType.JOB_TAG,
    state_type=StateType.DELTA_TOKEN,
    options={
        "job_name": args['JOB_NAME'],
        "logger": logger
    }
)

# Get connector options (including delta token if available)
key = "SAPODataNode"
connector_options = state_manager.get_connector_options(key)

# Use the connector options in your Glue job
df = glueContext.create_dynamic_frame.from_options(
    connection_type="SAPOData",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "entityName",
        "ENABLE_CDC": "true",
        **connector_options
    }
)

# Process your data here...

# Update the state after processing
state_manager.update_state(key, sapodata_df.toDF())
```

In both examples, the state management script handles the complexities of storing the state(either delta token or timestamp) between job runs. It automatically retrieves the last know state when getting connector options and updates the state after processing, ensuring the each job run only processes new or changed data.

Partitioning for Non ODP entities

In Apache Spark, partitioning refers to the way data is divided and distributed across the worker nodes in a cluster for parallel processing. Each partition is a logical chunk of data that can be processed independently by a task. Partitioning is a fundamental concept in Spark that directly impacts performance, scalability, and resource utilization. AWS Glue jobs use Spark's partitioning mechanism to divide the dataset into smaller chunks (partitions) that can be processed in parallel across the cluster's worker nodes. Note that partitioning is not applicable for ODP entities.

For more details, see AWS Glue Spark and PySpark jobs.

Prerequisites

An SAP OData’s Object you would like to read from. You will need the object/EntitySet name for example, `/sap/opu/odata/sap/API_SALES_ORDER_SRV/A_SalesOrder` .

Example

```
sapodata_read = glueContext.create_dynamic_frame.from_options(
    connection_type="SAPOData",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "/sap/opu/odata/sap/API_SALES_ORDER_SRV/A_SalesOrder"
    }, transformation_ctx=key)
```

Partitioning Queries

Field Based partitioning

You can provide the additional Spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, and `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently. Integer, Date and DateTime fields support field-based partitioning in the SAP OData connector.

- `PARTITION_FIELD`: the name of the field to be used to partition the query.
- `LOWER_BOUND`: an inclusive lower bound value of the chosen partition field.

  For any field whose data type is DateTime, the Spark timestamp format used in Spark SQL queries is accepted.

  Examples of valid values: `"2000-01-01T00:00:00.000Z"`
- `UPPER_BOUND`: an exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: number of partitions.
- `PARTITION_BY`: the type partitioning to be performed, `FIELD` to be passed in case of Field based partitioning.

Example

```
sapodata= glueContext.create_dynamic_frame.from_options(
    connection_type="sapodata",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "/sap/opu/odata/sap/SEPM_HCM_SCENARIO_SRV/EmployeeSet",
        "PARTITION_FIELD": "validStartDate"
        "LOWER_BOUND": "2000-01-01T00:00:00.000Z"
        "UPPER_BOUND": "2020-01-01T00:00:00.000Z"
        "NUM_PARTITIONS": "10",
        "PARTITION_BY": "FIELD"
    }, transformation_ctx=key)
```

Record Based partitioning

The original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

Record-based partitioning is only supported for non-ODP entities, as pagination in ODP entities is supported through the next token/skip token.

- `PARTITION_BY`: the type partitioning to be performed. `COUNT` is to be passed in case of record-based partitioning.

Example

```
sapodata= glueContext.create_dynamic_frame.from_options(
    connection_type="sapodata",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "/sap/opu/odata/sap/SEPM_HCM_SCENARIO_SRV/EmployeeSet",
        "NUM_PARTITIONS": "10",
        "PARTITION_BY": "COUNT"
    }, transformation_ctx=key)
```

Limitations / Callouts

- ODP entities are not compatible with Record Based Partitioning since pagination is handled using skip token/delta token. Consequently, for Record Based Partitioning, the default value for maxConcurrency is set to "null" irrespective of the user input.

- When both limit and partition is applied, the limit takes precedence over partitioning.

SAP OData connection options

The following are connection options for SAP OData:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your object in SAP OData.

  For example: /sap/opu/odata/sap/API_SALES_ORDER_SRV/A_SalesOrder
- `API_VERSION`(String) - (Optional) Used for Read. SAP OData Rest API version you want to use. Example: 2.0.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.

  For example: SalesOrder
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.

  For example: `SalesOrder = "10"`
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.

  For example: `SELECT * FROM /sap/opu/odata/sap/API_SALES_ORDER_SRV/A_SalesOrder`
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.

  For example: `ValidStartDate`
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.

  For example: `"2000-01-01T00:00:00.000Z"`
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field.

  For example: `"2024-01-01T00:00:00.000Z"`
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.
- `INSTANCE_URL`(String) - The SAP instance application host URL.

For example: `https://example-externaldata.sierra.aws.dev`

- `SERVICE_PATH`(String) - The SAP instance application service path.

  For example: `/sap/opu/odata/iwfnd/catalogservice;v=2`
- `CLIENT_NUMBER`(String) - The SAP instance application client number.

  For example: 100
- `PORT_NUMBER`(String) - Default: The SAP instance application port number.

  For example: 443
- `LOGON_LANGUAGE`(String) - The SAP instance application logon language.

  For example: `EN`
- `ENABLE_CDC`(String) - Defines whether to run a job with CDC enabled, that is, with track changes.

  For example: `True/False`
- `DELTA_TOKEN`(String) - Runs an incremental data pull based on the valid Delta Token supplied.

  For example: `D20241107043437_000463000`
- `PAGE_SIZE`(Integer) - Defines the page size for querying the records. The default page size is 50,000. When a page size is specified, SAP returns only the defined number of records per API call, rather than the entire dataset. The connector will still provide the total number of records and handle pagination using your specified page size. If you require a larger page size, you can choose any value up to 500,000, which is the maximum allowed. Any specified page size exceeding 500,000 will be ignored. Instead, the system will use the maximum allowed page size. You can specify the page size in the AWS Glue Studio UI by adding a connection option `PAGE_SIZE` with your desired value.

  For example: `20000`

SAP OData entity and field details

Entity	Data type	Supported operators
Tables (dynamic entities)	String	=, !=, >, >=, <, <=, BETWEEN, LIKE
	Integer	=, !=, >, >=, <, <=, BETWEEN, LIKE
	Long	=, !=, >, >=, <, <=, BETWEEN, LIKE
	Double	=, !=, >, >=, <, <=, BETWEEN, LIKE
	Date	=, !=, >, >=, <, <=, BETWEEN, LIKE
	DateTime	=, !=, >, >=, <, <=, BETWEEN, LIKE
	Boolean	=, !=
	Struct	=, !=, >, >=, <, <=, BETWEEN, LIKE

Connecting to SendGrid

SendGrid is a customer communication platform for transactional and marketing emails.

- SendGrid connector helps in creating and managing contact lists and creating email marketing campaigns.
- SendGrid allows online businesses, non-profits, and other online entities to create and send marketing emails to large audiences and monitor engagement with those emails.

Topics

- AWS Glue support for SendGrid

- Policies containing the API operations for creating and using connections
- Configuring SendGrid
- Configuring SendGrid connections
- Reading from SendGrid entities
- SendGrid connection options
- SendGrid limitations

AWS Glue support for SendGrid

AWS Glue supports SendGrid as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from SendGrid.

Supported as a target?

No.

Supported SendGrid API versions

The following SendGrid API versions are supported:

- v3

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
                "glue:DescribeConnectionType",
                "glue:RefreshOAuth2Tokens",
                "glue:ListEntities",
                "glue:DescribeEntity"
            ],
            "Resource": "*"
        }
    ]
}
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring SendGrid

Before you can use AWS Glue to transfer data from SendGrid, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- You have a SendGrid account with an API key.
- Your SendGrid account has API access with a valid license.

If you meet these requirements, you're ready to connect AWS Glue to your SendGrid account. For typical connections, you don't need do anything else in SendGrid.

Configuring SendGrid connections

SendGrid supports custom authentication.

For public SendGrid documentation on generating the required API keys for custom authentication, see Authentication.

To configure a SendGrid connection:

1. In AWS Secrets Manager, create a secret with the following details:
   a. For the customer managed connected app, the Secret should contain the connected app Consumer Secret with *`api_key`* as key.
   b. Note: you must create a secret for your connections in AWS Glue.

1. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Connection type**, select SendGrid.
   b. Provide the `INSTANCE_URL` of the SendGrid instance you want to connect to.
   c. Select the AWS IAM role which AWS Glue can assume and has permissions for following actions:

      JSON

      ```
      {
        "Version": "2012-10-17",
        "Statement": [
          {
            "Effect": "Allow",
            "Action": [
              "secretsmanager:DescribeSecret",
              "secretsmanager:GetSecretValue",
              "secretsmanager:PutSecretValue",
              "ec2:CreateNetworkInterface",
              "ec2:DescribeNetworkInterfaces",
              "ec2:DeleteNetworkInterface"
            ],
            "Resource": "*"
          }
        ]
      }
      ```

   d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.

   e. Select the network options if you want to use your network.

2. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from SendGrid entities

Prerequisite

A SendGrid object you would like to read from. You will need the object name such as `lists`, `singlesends` or `segments`.

Supported entities for source:

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Lists	No	Yes	No	Yes	No
Single Sends	Yes	Yes	No	Yes	No
Marketing Campaign Stats-Automations	Yes	Yes	No	Yes	No
Marketing Campaign Stats-Single Sends	Yes	Yes	No	Yes	No
Segments	Yes	No	No	Yes	No
Contacts	Yes	No	No	Yes	No
Category	No	No	No	Yes	No
Stats	Yes	No	No	Yes	No
Unsubscribe Groups	Yes	No	No	Yes	No

Example:

```
sendgrid_read = glueContext.create_dynamic_frame.from_options(
    connection_type="sendgrid",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "lists",
        "API_VERSION": "v3",
        "INSTANCE_URL": "instanceUrl"
    }
```

SendGrid entity and field details:

Entities with static metadata:

Entity	Field	Data type	Supported operators
Lists	id	String	N/A
	name	String	N/A
	contact_count	Integer	N/A
	_metadata	Struct	N/A
Single Sends	id	String	N/A
	name	String	EQUAL_TO
	abtest	Struct	N/A
	status	String	EQUAL_TO
	categories	List	EQUAL_TO
	send_at	String	N/A
	is_abtest	Boolean	N/A
	updated_at	String	N/A

Entity	Field	Data type	Supported operators	
	created_at	String	N/A	
	channels	List	N/A	
Marketing Campaign Stats-Automations	id	String	N/A	
	aggregation	String	N/A	
	step_id	String	N/A	
	stats	Struct	N/A	
	automation_ids	List	EQUAL_TO	
Marketing Campaign Stats-Singlesends	id	String	N/A	
	ab_variation	String	N/A	
	ab_phase	String	N/A	
	aggregation	String	N/A	
	stats	Struct	N/A	
	singlesend_ids	List	EQUAL_TO	
Segments	id	String	N/A	
	name	String	N/A	
	query_version	String	N/A	
	contacts_count	Integer	N/A	
	sample_updated_at	String	N/A	
	next_sample_update	String	N/A	

Entity	Field	Data type	Supported operators	
	created_at	String	N/A	
	updated_at	String	N/A	
	parent_list_id	String	N/A	
	status	Struct	N/A	
	parent_list_ids	String	EQUAL_TO	
	no_parent _list_id	Boolean	EQUAL_TO	
Contacts	id	String	N/A	
	first_name	String	N/A	
	last_name	String	N/A	
	unique_name	String	N/A	
	email	String	N/A	
	alternate_emails	List	N/A	
	address_line_1	String	N/A	
	address_line_2	String	N/A	
	city	String	N/A	
	state_pro vince_region	String	N/A	
	country	String	N/A	
	postal_code	String	N/A	
	phone_number	String	N/A	

Entity	Field	Data type	Supported operators	
	whatsapp	String	N/A	
	line	String	N/A	
	facebook	String	N/A	
	list_ids	List	N/A	
	custom_fields	Struct	N/A	
	created_at	String	N/A	
	updated_at	String	N/A	
	_metadata	Struct	N/A	
	event_timestamp	DateTime	BETWEEN	
Category	categories	List	N/A	
Stats	date	String	N/A	
	stats	List	N/A	
	start_date	DateTime	EQUAL_TO, BETWEEN	
	aggregated_by	String	EQUAL_TO	
Unsubscribe Groups	id	Integer	EQUAL_TO	
	name	String	N/A	
	description	String	N/A	
	last_email_sent_at	Integer	N/A	

Entity	Field	Data type	Supported operators
	is_default	Boolean	N/A
	unsubscribes	Integer	N/A

Note

Struct and List data types are converted to String data type, and DateTime data type is converted to Timestamp in the response of the connectors.

Partitioning queries

SendGrid doesn't support filter-based partitioning or record-based partitioning.

SendGrid connection options

The following are connection options for SendGrid:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your object in SendGrid.
- `API_VERSION`(String) - (Required) Used for Read. SendGrid Rest API version you want to use.
- `INSTANCE_URL`(String) - (Required) Used for Read. A valid SendGrid instance URL.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.

SendGrid limitations

The following are limitations or notes for SendGrid:

- Incremental pull is only supported by the Stats entity on the `start_date` field and by the Contact entity on the `event_timestamp` field.

- Pagination is only supported by the Marketing Campaign Stats (Automations), Marketing Campaign Stats (Single Sends), Single Sends, and Lists entities.
- For the Stats entity, `start_date` is a mandatory filter parameter.
- An API key with Restricted Access can't support read access for the Email API and Stats entities. Use an API key with Full Access. For more information, see API Overview.

Connecting to ServiceNow

ServiceNow is a cloud-based SaaS platform for automating IT management workflows. The ServiceNow platform easily integrates with other tools, letting users manage projects, teams and customer interactions using a variety of apps and plugins. As a ServiceNow user you can connect AWS Glue to your ServiceNow account. Then, you can use ServiceNow as a data source in your ETL jobs. Run these jobs to transfer data between ServiceNow and AWS services or other supported applications.

Topics

- AWS Glue support for ServiceNow
- Policies containing the API operations for creating and using connections
- Configuring ServiceNow
- Configuring ServiceNow connections
- Reading from ServiceNow entities
- ServiceNow connection options
- Limitations and notes for ServiceNow connector

AWS Glue support for ServiceNow

AWS Glue supports ServiceNow as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from ServiceNow.

Supported as a target?

No.

Supported ServiceNow API versions

The following ServiceNow API versions are supported:

- v2

For entity support per version specific, see Supported entities for Source.

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention

for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring ServiceNow

Before you can use AWS Glue to transfer data from ServiceNow, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- You have a ServiceNow account with email and password. For more information, see Creating a ServiceNow account.
- Your ServiceNow account is enabled for API access. All use of the ServiceNow API is available at no additional cost.

If you meet these requirements, you're ready to connect AWS Glue to your ServiceNow account.

Creating a ServiceNow account

To create a ServiceNow account:

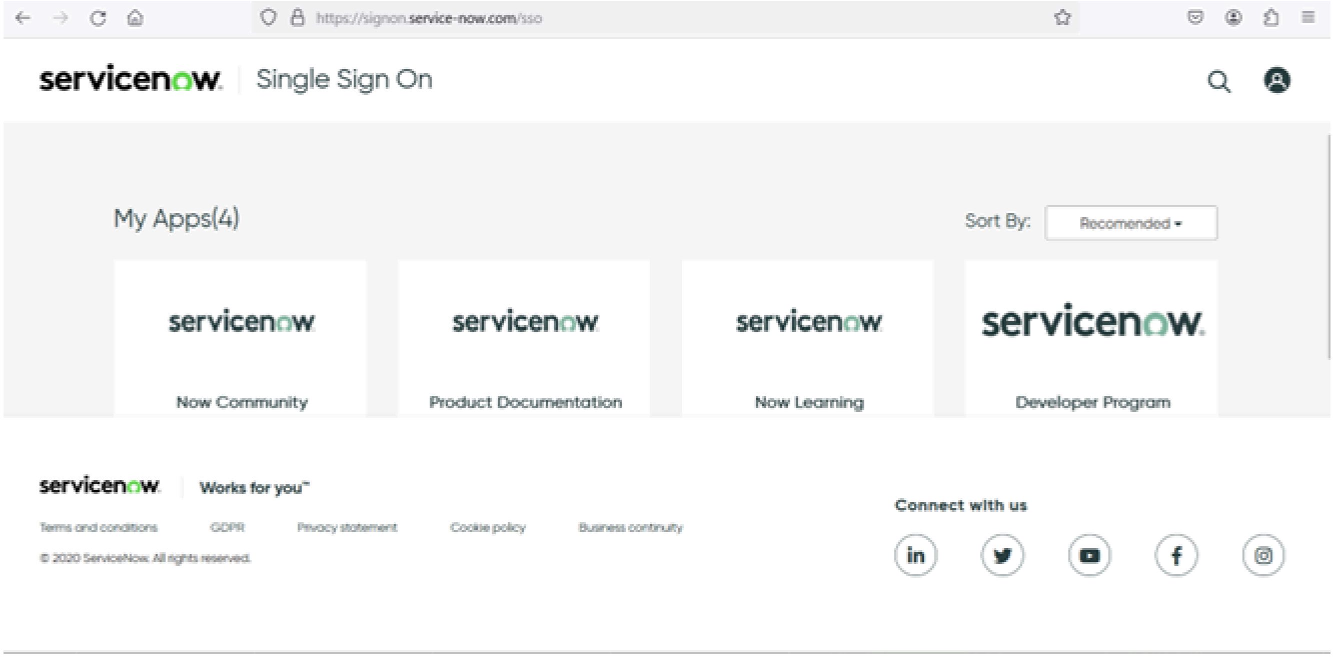
1. Navigate to the sign up page on servicenow.com, enter your details, and click **Continue**.
2. When you receive a verification code in your registered mail, enter that code and choose **Verify**.
3. Set up multi-factor authentication or skip doing so.

Your account is created and ServiceNow displays your profile.

Creating a ServiceNow developer instance

Request a developer instance after logging in to ServiceNow.

1. At the ServiceNow login page, enter your account credentials.
2. Choose the **ServiceNow Developer Program**.


https://signon.service-now.com/sso
servicenow Single Sign On
My Apps(4)
Sort By:
Recomended
Now Community
Product Documentation
Now Learning
Developer Program
Works for you™
Terms and conditions
GDPR
Privacy statement
Cookie policy
Business continuity
© 2020 ServiceNow. All rights reserved.
Connect with us

3. Choose **Request Instance** in the top right.
4. Enter your job responsibilities. Indicate your agreement to the terms of use, and choose **Finish setup**.
5. Once the instance is created, note your instance URL and credentials.

Retrieving BasicAuth credentials

To retrieve Basic Auth credentials for a free account:

1. At the ServiceNow login page, enter your account credentials.
2. On the home page choose the edit profile section (top right corner) and choose **Manage Instance Password**.
3. Retrieve the login credentials such as username, password, and instance URL.

> **Note**
>
> If MFA is enabled for the account, append the MFA token to the end of the user's password in the basic auth: <username>:<password><MFA Token>

For more information, see Building applications in the ServiceNow documentation.

Creating OAuth 2.0 credentials

To use OAuth2.0 in ServiceNow connector, you need to create an inbound client) to generate the Client ID and Client Secret:

1. At the ServiceNow login page, enter your account credentials.
2. On the home page choose **Start Building**.
3. On the App Engine Studio page, search for **Application Registry**.
4. Choose **New** in the top right.
5. Choose the **Create an OAuth API endpoint for external clients** option.
6. Make any required changes to the OAuth configuration and choose **Update**.

   Example for Redirect URL: https://us-east-1.console.aws.amazon.com/gluestudio/oauth
7. Select the newly created OAuth client app to retrieve the Client ID and Client Secret.
8. Store the Client ID and Client Secret for further processing.

To configure OAuth in a non-production developer account:

1. Create an authentication profile using the Create an authentication profile topic in the ServiceNow documentation.
2. In the Authentication Profile for OAuth, select **Type** as OAuth and select the above-created inbound client to set the **OAuth Entity**.
3. If there are multiple clients, then you need to create multiple authentication profiles to set the required OAuth entity in the authentication profile.
4. If not configured, create a REST API access policy, to give access to the TABLE API. See Create REST API access policy.

Configuring ServiceNow connections

The grant type determines how AWS Glue communicates with ServiceNow to request access to your data. Your choice affects the requirements that you must meet before you create the connection. ServiceNow supports only the AUTHORIZATION_CODE grant type for OAuth 2.0.

- This grant type is considered "three-legged" OAuth as it relies on redirecting users to a third-party authorization server to authenticate the user. It is used when creating connections via the

AWS Glue console. The AWS Glue console will redirect the user to ServiceNow where the user must login and allow AWS Glue the requested permissions to access their ServiceNow instance.

- Users may still opt to create their own connected app in ServiceNow and provide their own client id and client secret when creating connections through the AWS Glue console. In this scenario, they will still be redirected to ServiceNow to login and authorize AWS Glue to access their resources.
- This grant type results in a refresh token and access token. The access token is short lived, and may be refreshed automatically without user interaction using the refresh token.
- For public ServiceNow documentation on creating a connected app for Authorization Code OAuth flow, see Set up OAuth.

To configure a ServiceNow connection:

1. In AWS Secrets Manager, create a secret with the following details:
   a. For basic authentication, the Secret should contain the connected app Consumer Secret with `USERNAME` and `PASSWORD` as key.
   b. For an authorization code grant type, the Secret should contain the connected app Consumer Secret with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as key.
   c. Note: You must create a secret per connection in AWS Glue.
2. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Connection type**, select ServiceNow.
   b. Provide the INSTANCE_URL of the ServiceNow instance you want to connect to.
   c. Select the AWS IAM role which AWS Glue can assume and has permissions for following actions:

      JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
          "ec2:CreateNetworkInterface",
          "ec2:DescribeNetworkInterfaces",
          "ec2:DeleteNetworkInterface"
        ],
        "Resource": "*"
      }
    ]
  }
```

   d. Select the **Authentication Type** you want to use for this connection in AWS Glue.

      i. Basic Auth: this auth type works well for automation scenarios as it allows to use username and password up front with the permissions of a particular user in the ServiceNow instance. AWS Glue is able to use the username and password to authenticate ServiceNow APIs. Enter the following inputs only in case of Basic Auth: `Username` and `Password`.

      ii. OAuth2: enter the following inputs only in case of OAuth2: `ClientId`, `ClientSecret`, `Authorization URL`, `Authorization Token URL`.

   e. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.

   f. Select the network options if you want to use your network.

3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from ServiceNow entities

Prerequisite

A ServiceNow Tables object you would like to read from. You will need the object name such as pa_bucket or incident.

Example:

```
servicenow_read = glueContext.create_dynamic_frame.from_options(
    connection_type="servicenow",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "pa_buckets",
        "API_VERSION": "v2"
        "instanceUrl": "https://<instance-name>.service-now.com"
    }
)
```

ServiceNow entity and field details:

For the following entities, ServiceNow provides endpoints to fetch metadata dynamically, so that operator support is captured at the datatype level for each entity.

Entity	Data type	Supported operators
Tables (dynamic entities)	Integer	=, !=, <, <=, >, >=, BETWEEN
	BigDecimal	=, !=, <, <=, >, >=, BETWEEN
	Float	=, !=, <, <=, >, >=, BETWEEN
	Long	=, !=, <, <=, >, >=, BETWEEN
	Date	=, !=, <, <=, >, >=, BETWEEN
	DateTime	=, !=, <, <=, >, >=, BETWEEN
	Boolean	=, !=
	String	=, !=, <, <=, >, >=, BETWEEN, LIKE
	Struct	N/A

> **Note**
>
> The Struct data type is converted to a String data type in the response of the connector.

> **Note**
>
> `DML_STATUS` is an additional user-defined attribute used for tracking CREATED/UPDATED records.

Partitioning queries

Field base partitioning:

You can provide the additional Spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, and `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

Entity name	Partitioning fields	Data type
Dynamic Entity	sys_mod_count	Integer
	sys_created_on, sys_updated_on	DateTime

- `PARTITION_FIELD`: the name of the field to be used to partition the query.
- `LOWER_BOUND`: an **inclusive** lower bound value of the chosen partition field.

  For the Datetime field, we accept the Spark timestamp format used in SPark SQL queries.

  Examples of valid value:

  ```
  "2024-01-30T06:47:51.000Z"
  ```
- `UPPER_BOUND`: an **exclusive** upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: the number of partitions.

The following table describes the entity partitioning field support details:

Example:

```
servicenow_read = glueContext.create_dynamic_frame.from_options(
    connection_type="servicenow",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "pa_buckets",
        "API_VERSION": "v2",
        "instanceUrl": "https://<instance-name>.service-now.com"
        "PARTITION_FIELD": "sys_created_on"
        "LOWER_BOUND": "2024-01-30T06:47:51.000Z"
        "UPPER_BOUND": "2024-06-30T06:47:51.000Z"
        "NUM_PARTITIONS": "10"
```

```
    }
```

Record-based partitioning:

You can provide the additional Spark option `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With this parameter, the original query is split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

In record-based partitioning, the total number of records present is queried from the ServiceNow API, and it is divided by `NUM_PARTITIONS` number provided. The resulting number of records are then concurrently fetched by each sub-query.

- `NUM_PARTITIONS`: the number of partitions.

Example:

```
servicenow_read = glueContext.create_dynamic_frame.from_options(
    connection_type="servicenow",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "pa_buckets",
        "API_VERSION": "v2",
        "instanceUrl": "https://<instance-name>.service-now.com"
        "NUM_PARTITIONS": "2"
    }
```

ServiceNow connection options

The following are connection options for ServiceNow:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your object in ServiceNow.
- `API_VERSION`(String) - (Required) Used for Read. ServiceNow Rest API version you want to use. For example: v1,v2,v3,v4.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.

- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field. For example: 2024-01-30T06:47:51.000Z.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field. For example: 2024-06-30T06:47:51.000Z.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read. For example: 10.
- `INSTANCE_URL`(String) - (Required) A valid ServiceNow instance URL with format https://<instance-name>.service-now.com.
- `PAGE_SIZE`(Integer) - Defines the page size for querying the records. The default page size is 1,000. When a page size is specified, ServiceNow returns only the defined number of records per API call, rather than the entire dataset. The connector will still provide the total number of records and handle pagination using your specified page size. If you require a larger page size, you can choose any value up to 10,000, which is the maximum allowed. Any specified page size exceeding 10,000 will be ignored. Instead, the system will use the maximum allowed page size. You can specify the page size in the AWS Glue Studio UI by adding a connection option `PAGE_SIZE` with your desired value. For example: 5000.

Limitations and notes for ServiceNow connector

The following are limitations or notes for the ServiceNow connector:

- As per SaaS documentation, `sys_created_on`, `sys_updated_on`, and `sys_mod_count` are system generated fields. The connector relies on SaaS APIs to provide these fields in the response body.
  - If SaaS doesn't generate these fields for any entity, filter based partitioning cannot be supported.
- If SaaS APIs don't return `sys_created_on` and `sys_updated_on` fields in the response, `DML_STATUS` cannot be calculated.

Connecting to Slack in AWS Glue Studio

Slack is an enterprise communications app that lets users send messages and attachments through various public and private channels. If you're a Slack user, you can connect AWS Glue to your Slack account. Then, you can use Slack as a data source in your ETL Jobs. Run these jobs to transfer data between Slack and AWS services or other supported applications.

Topics

- AWS Glue support for Slack
- Policies containing the API operations for creating and using connections
- Configuring Slack
- Configuring Slack connections
- Reading from Slack entities
- Slack connection options
- Limitations
- Creating a new Slack account and configuring the client app

AWS Glue support for Slack

AWS Glue supports Slack as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Slack.

Supported as a target?

No.

Supported Slack API versions

Slack API v2.

Policies containing the API operations for creating and using connections

The following sample policy describes the required IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
        "Effect": "Allow",
        "Action": [
          "glue:ListConnectionTypes",
          "glue:DescribeConnectionType",
          "glue:RefreshOAuth2Tokens",
          "glue:ListEntities",
          "glue:DescribeEntity"
        ],
        "Resource": "*"
      }
    ]
  }
```

You can also use the following managed IAM policies to allow access:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, Amazon CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Slack

Before you can use AWS Glue to transfer data to or from Slack, you must meet these requirements:

Minimum requirements

- You must have a Slack account. For more information, see Creating a new Slack account and configuring the client app.

If you meet these requirements, you’re ready to connect AWS Glue to your Slack account.

Configuring Slack connections

Slack supports the `AUTHORIZATION_CODE` grant type for OAuth 2.

This grant type is considered “three-legged” OAuth as it relies on redirecting users to the third party authorization server to authenticate the user. It is used when creating connections via the AWS Glue Console. The AWS Glue Console will redirect the user to Slack where the user must login and allow AWS Glue the requested permissions to access their Slack instance.

Users may still opt to create their own connected app in Slack and provide their own client id and client secret when creating connections through the AWS Glue Console. In this scenario, they will still be redirected to Slack to login and authorize AWS Glue to access their resources.

This grant type results in a refresh token and access token. The access token expires after 1 hour of creation. A new access token can be fetched using the refresh token.

For more information on creating a connected app for Authorization Code OAuth flow, see Slack API .

To configure an Slack connection:

1. In AWS Secrets Manager, create a secret with the following details. You must create a secret for the connection in AWS Glue.
   a. For customer managed connected app – Secret should contain the connected app Consumer Secret with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as key.
2. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Connection type**, select Slack.
   b. Provide the Slack environment.
   c. Select the IAM role which AWS Glue can assume and has permissions for following actions:

      JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
                "ec2:DeleteNetworkInterface"
            ],
            "Resource": "*"
        }
    ]
}
```

d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.

e. Select the network options if you want to use your network.

3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Slack entities

Prerequisites

- A Slack object you would like to read from.

Supported entities

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
conversations	Yes	Yes	No	Yes	Yes

Example

```
slack_read = glueContext.create_dynamic_frame.from_options(
    connection_type="slack",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "conversations/C058W38R5J8"
    }
)
```

Slack entity and field details

Entity	Field	Data Type	Supported Operators
conversations	attachments	List	NA
conversations	bot_id	String	NA
conversations	blocks	List	NA
conversations	client_msg_id	String	NA
conversations	is_starred	Boolean	NA
conversations	last_read	String	NA
conversations	latest_reply	String	NA
conversations	reactions	List	NA
conversations	replies	List	NA
conversations	reply_count	Integer	NA
conversations	reply_users	List	NA
conversations	reply_users_count	Integer	NA
conversations	subscribed	Boolean	NA
conversations	subtype	String	NA
conversations	text	String	NA
conversations	team	String	NA
conversations	thread_ts	String	NA
conversations	ts	String	EQUAL_TO, BETWEEN, LESS_THAN , LESS_THAN

Entity	Field	Data Type	Supported Operators
			_OR_EQUAL _TO, GREATER_T HAN, GREATER_T HAN_OR_EQUAL_TO
conversations	type	String	NA
conversations	user	String	NA
conversations	inviter	String	NA
conversations	root	Struct	NA
conversations	is_locked	Boolean	NA
conversations	files	List	NA
conversations	room	Struct	NA
conversations	upload	Boolean	NA
conversations	display_as_bot	Boolean	NA
conversations	channel	String	NA
conversations	no_notifications	Boolean	NA
conversations	permalink	String	NA
conversations	pinned_to	List	NA
conversations	pinned_info	Struct	NA
conversations	edited	Struct	NA
conversations	app_id	String	NA
conversations	bot_profile	Struct	NA

Entity	Field	Data Type	Supported Operators
conversations	metadata	Struct	NA

Partitioning queries

Additional spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, `NUM_PARTITIONS` can be provided if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition query.
- `LOWER_BOUND`: an inclusive lower bound value of the chosen partition field.

  For date, we accept the Spark date format used in Spark SQL queries. Example of valid value: `"2024-07-01T00:00:00.000Z"`.
- `UPPER_BOUND`: an exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: number of partitions.

Entity wise partitioning field support details are captured in below table.

Entity Name	Partitioning Field	Data Type
conversations	ts	String

Example

```
slack_read = glueContext.create_dynamic_frame.from_options(
    connection_type="slack",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "conversations/C058W38R5J8",
        "PARTITION_FIELD": "ts"
        "LOWER_BOUND": "2022-12-01T00:00:00.000Z"
        "UPPER_BOUND": "2024-09-23T15:00:00.000Z"
        "NUM_PARTITIONS": "2"
    }
)
```

Slack connection options

The following are connection options for Slack:

- `ENTITY_NAME`(String) - (Required) Used for Read. Supported entity name. Example: `conversations/C058W38R5J8`.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Fields you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.

Limitations

The following are limitations for the Slack connector:

- Record based partitioning is not supported as connector does not provide any means to fetch the total number of records (messages) available in a given conversation.

Creating a new Slack account and configuring the client app

Creating a Slack account

1. Open the Slack home page to sign-up for an account.
2. Choose **SIGN UP WITH EMAIL ADDRESS**. Enter your email ID and choose **Continue**.
3. Enter the 6-character code sent to your email address, it will redirect you to create a workspace or to join an existing workspace.
4. Choose **Create a workspace** to create a new workspace. It will redirect you to answer a few questions as a part of the set-up process.

    - Name of company

- Your name
- To add colleagues by email
- What's your team working on? (This will be the channel name)

5. Fill in the input fields for these questions and continue. Your account is now ready to be used.

Creating a Slack developer app

1. Log in to your Slack account and sign into your Slack workspace.
2. From the workspace menu, select **Tools and settings** and then select **Manage apps**.
3. From the Slack App Directory menu, select **Build**.
4. On the **Your Apps** page, select **Create an App**.
5. On the **Create an app** page, select **From scratch**.
6. In the **Name app & choose workspace** dialog box that opens, add an App name and **Pick a workspace to deploy your app in**. Then select **Create App**.
7. Note down your Client Id and Secret displayed in App Credentials
8. On the OAuth & Permissions sidebar, go to Scopes and choose **Add an OAuth Scope**. You can add the redirect URLs to your app for configuration to automatically generate the 'Add to Slack' button or to distribute your app. Scroll up to the Redirect URLs section and choose **Add New Redirect URL** and save.
9. Then, scroll to OAuth Tokens for Your Workspace section, and choose **Install to Workspace**.
10. On the dialog box that opens up informing you that the app that you created is requesting permission to access the Slack workspace you wanted to connect it to, select **Allow**.
11. On successful completion, the console will display a OAuth Tokens for Your Workspace screen.
12. From the OAuth Tokens for Your Workspace screen, copy and save the OAuth token you will use to connect to AWS Glue
13. Next, you retrieve your Slack team ID. From the Slack workspace menu, select **Tools and settings** and then select **Manage apps**. You'll find your team ID in the URL of the page that opens.
14. To publicly distribute your app, you can activate by heading over to the **Manage Distribution** button on the sidebar. Scroll down to the Share Your App with Other Workspaces section and choose **Remove Hard Coded Information**. Provide consent and choose **Active Public Distribution**.

15. Your app is now publicly distributed. To access the entity APIs, the app needs to be added to every workspace channel the user wants to access from.
16. Sign into your slack account and open the workspace whose channel needs to be accessed.
17. In the workspace, open the channel for which the app wants to access and choose the channel title. Select the **Integrations** tab from the pop-up and add the app. In this way, the app is integrated with the channel to have access to its API.

    The OAuth 2.0 client ID must have one or more authorized redirect URLs. Redirect URLs have the following format:

    > **Note**
    >
    > Appflow redirect URLs are subject to change post redirect URLs for AWS Glue platform are available. Client ID and Client Secret are from the settings for your OAuth 2.0 client ID.

Redirect URL can be one of the following	
Redirect URL for Gamma environment	
https://us-west-2.console.aws.amazon.com/appflow/oauth	https://us-east-1.awsc-integ.aws.amazon.com/appflow/oauth
https://us-east-2.console.aws.amazon.com/appflow/oauth	
https://us-west-1.console.aws.amazon.com/appflow/oauth	
https://ap-south-1.console.aws.amazon.com/appflow/oauth	
https://ap-southeast-1.console.aws.amazon.com/appflow/oauth	
https://ap-southeast-2.console.aws.amazon.com/appflow/oauth	

Redirect URL can be one of the following	
Redirect URL for Gamma environment	
https://ap-northeast-1.console.aws.amazon.com/appflow/oauth	
https://ap-northeast-2.console.aws.amazon.com/appflow/oauth	
https://ca-central-1.console.aws.amazon.com/appflow/oauth	
https://eu-central-1.console.aws.amazon.com/appflow/oauth	
https://eu.west-1.console.aws.amazon.com/appflow/oauth	
https://eu-west-2.console.aws.amazon.com/appflow/oauth	
https://eu-west-3.console.aws.amazon.com/appflow/oauth	
https://sa-east-1.console.aws.amazon.com/appflow/oauth	
https://us-west-2.awsc-integ.aws.amazon.com/appflow/oauth	
https://af-south-1.console.aws.amazon.com/appflow/oauth	

Connecting to Smartsheet

Smartsheet is a work management and collaboration SaaS product. Fundamentally, Smartsheet allows users to use spreadsheet-like objects to create, store, and utilize business data.

Topics

- AWS Glue support for Smartsheet
- Policies containing the API operations for creating and using connections
- Configuring Smartsheet
- Configuring Smartsheet connections
- Reading from Smartsheet entities
- Smartsheet connection options
- Creating an Smartsheet account
- Limitations

AWS Glue support for Smartsheet

AWS Glue supports Smartsheet as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Smartsheet.

Supported as a target?

No.

Supported Smartsheet API versions

v2.0

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the preceding method, alternatively, use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Smartsheet

Before you can use AWS Glue to transfer from Smartsheet, you must meet the following requirements:

Minimum requirements

- You have an Smartsheet account with email and password. For more information about creating an account, see Creating a Smartsheet account.
- Your Smartsheet account has API access with valid license.
- Your Smartsheet account has **Pro pricing** plan for `Sheets` entity and Enterprise pricing plan with Event Reporting Add-On for `Events` entity.

If you meet these requirements, you're ready to connect AWS Glue to your Smartsheet account. For typical connections, you don't need do anything else in Smartsheet.

Configuring Smartsheet connections

Smartsheet supports `AUTHORIZATION_CODE` grant type for OAuth2.

This grant type is considered “three-legged” `OAuth` as it relies on redirecting users to the third-party authorization server to authenticate the user. Users may opt to create their own connected app in Smartsheet and provide their own client ID and client secret when creating connections through the AWS Glue console. In this scenario, they will still be redirected to Smartsheet to login and authorize AWS Glue to access their resources.

This grant type results in a refresh token and access token. The access token is short lived, and may be refreshed automatically without user interaction using the refresh token.

For public Smartsheet documentation on creating a connected app for AUTHORIZATION_CODE OAuth flow, see Smartsheet APIs .

To configure a Smartsheet connection:

1. In AWS Secrets Manager, create a secret with the following details:

   For customer managed connected app – Secret should contain the connected app Consumer Secret with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as key.

   > **Note**
   >
   > It is a must to create a secret per connection in AWS Glue.

2. In the AWS Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Connection type**, select Smartsheet.
   b. Provide the `instanceUrl` of the Smartsheet you want to connect to.
   c. Select the IAM role for which AWS Glue can assume and has permissions for following actions:

      JSON

      ```
      {
        "Version": "2012-10-17",
      ```

```
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.

e. Select the network options if you want to use your network.

3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Smartsheet entities

Prerequisites

A `Smartsheet` Object you would like to read from. Refer the supported entities table below to check the available entities.

Supported entities

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
List Sheet	Yes	Yes	No	Yes	No
Row Metadata	Yes	Yes	No	Yes	No
Sheet Metadata	No	No	No	Yes	No

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Sheet Data	Yes	Yes	Yes	Yes	No
Event	Yes	Yes	No	Yes	No

Example

```
Smartsheet_read = glueContext.create_dynamic_frame.from_options(
    connection_type="smartsheet",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "list-sheets",
        "API_VERSION": "2.0",
        "INSTANCE_URL": "https://api.smartsheet.com"
    })
```

Smartsheet entity and field details

Entity	Field	Data Type	Supported Operators
List Sheets	id	Long	NA
	accessLevel	String	NA
	createdAt	DateTime	NA
	modifiedAt	DateTime	NA
	name	String	NA
	permalink	String	NA
	modifiedSince	DateTime	>=
	version	Integer	NA
	source	Struct	NA

Entity	Field	Data Type	Supported Operators
Row Metadata	id	Long	NA
	sheetId	Long	NA
	accessLevel	String	NA
	attachments	List	NA
	columns	List	NA
	conditionalFormat	String	NA
	createdAt	DateTime	NA
	createdBy	Struct	NA
	discussions	List	NA
	proofs	Struct	NA
	expanded	Boolean	NA
	filteredOut	Boolean	NA
	format	String	NA
	inCriticalPath	Boolean	NA
	locked	Boolean	NA
	lockedForUser	Boolean	NA
	modifiedAt	DateTime	NA
	modifiedBy	Struct	NA
	permalink	String	NA
	rowNumber	Integer	NA

Entity	Field	Data Type	Supported Operators
	version	Integer	NA
	totalRowCount	Integer	NA
	rowsModifiedSince	DateTime	>
	filterId	Long	“="
	siblingId	Long	NA
	parentId	Long	NA
Sheet metadata	id	Long	NA
	fromId	Long	NA
	ownerId	Long	NA
	accessLevel	String	NA
	attachments	List	NA
	columns	List	NA
	createdAt	DateTime	NA
	crossSheetReferences	List	NA
	dependenciesEnabled	Boolean	NA
	discussions	List	NA
	effectiveAttachmentOptions	List	NA
	favorite	Boolean	NA
	ganttEnabled	Boolean	NA

Entity	Field	Data Type	Supported Operators
	hasSummaryFields	Boolean	NA
	modifiedAt	DateTime	NA
	name	String	NA
	owner	String	NA
	permalink	String	NA
	projectSettings	Struct	NA
	readOnly	Boolean	NA
	resourceManagementEnabled	Boolean	NA
	showParentRowsForFilters	Boolean	NA
	source	Struct	NA
	summary	Struct	NA
	totalRowCount	Integer	NA
	userPermissions	Struct	NA
	userSettings	Struct	NA
	version	Integer	NA
	workspace	Struct	NA
	filters	List	NA
	ganttConfig	Struct	NA

Entity	Field	Data Type	Supported Operators
	resourceManagementType	String	NA
	cellImageUploadEnabled	Boolean	NA
	isMultiPicklistEnabled	Boolean	NA
Events	eventId	String	NA
	objectType	String	NA
	action	String	NA
	objectId	Long	NA
	eventTimestamp	DateTime	NA
	userId	Long	NA
	requestUserId	Long	NA
	accessTokenName	String	NA
	source	String	NA
	additionalDetails	Struct	NA
	since	DateTime	>=

Entities with dynamic metadata:

For the following entity, Smartsheet provides an endpoint to fetch metadata dynamically, allowing operator support to be captured at the datatype level.

Entity	Data Type	Supported Operators
Sheet Data	String	NA
Sheet Data	Long	"="
Sheet Data	Integer	NA
Sheet Data	DateTime	>

Example

```
Smartsheet_read = glueContext.create_dynamic_frame.from_options(
    connection_type="smartsheet",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "list-sheets",
        "API_VERSION": "2.0",
        "INSTANCE_URL": "https://api.smartsheet.com"
    }
```

Smartsheet connection options

The following are connection options for Smartsheet:

- `ENTITY_NAME`(String) – (Required) Used for Read/Write. The name of your Object in Smartsheet.
- `API_VERSION`(String) – (Required) Used for Read/Write. Smartsheet Rest API version you want to use. For example: v2.0.
- `INSTANCE_URL`(String) – (Required) Used for Read. Smartsheet instance URL.
- `SELECTED_FIELDS`(List<String>) – Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) – Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) – Default: empty. Used for Read. Full Spark SQL query.

Creating an Smartsheet account

1. Sign up for a Smartsheet account by accessing Smartsheet sing-up page.
2. Choose **Create one** to create a new account, or sign in using your registered Google, Microsoft, or Apple account.
3.
4.
5. Open the confirmation email from Smartsheet, and choose the confirmation link to verify your account.

   You will be subscribed to the Trial Plan by default.
6. In the bottom-left corner, choose the **Account** icon and choose **Add Licenses/Upgrade** to upgrade your pricing plan.

   > **Note**
   >
   > This is required for accessing **Event Reporting**, which is an add-on in the **Enterprise** plan.
7. Under the **Enterprise** plan, choose **Contact Us** to request an account upgrade from the support team.
8. In the support request form, provide the required details and your requirements to upgrade the plan.

   This completes the upgrade to **Enterprise** plan.

Creating `OAuth2.0` credentials

1. After upgrading your account's pricing plan to get access to the **Developer Tools**, access Smartsheet developers.

   You will receive an activation email.
2. Open an activation email from Smartsheet, and choose the activation link to activate developer tools on your account.

   Developers tool allows you to create the app.
3. Open the home page of your Smartsheet account and choose **Account** to check for access.

4. Choose **Developer Tools** from the services list, and enter the **Developer Profile** details.
5. Choose **Create New App**.
6. Enter the following details into the app registration form:
    - **Name –** Name of the app.
    - **Description –** Description of the app.
    - **URL –** URL that allows you to launch your app or the URL of the landing page.
    - **Contact/support –** Contact information for the support team.
    - **Redirect URL –** URL (also known as a callback URL) within your application that will receive the OAuth 2.0 credentials.
7. Choose **Save**.

    Smartsheet assigns a client ID and client secret to your app. Record these values for the next steps. You can also look them up again later in the **Developer Tools** section.

Limitations

Smartsheet doesn't support field-based or record-based partitioning.

Connecting to Snapchat Ads in AWS Glue Studio

Snapchat is a multimedia instant messaging app and service developed by Snap Inc., originally Snapchat Inc. One of the principal features of Snapchat is that pictures and messages are usually only available for a brief time before they become inaccessible to their recipients. Snapchat Marketing are posts for which businesses can pay to serve to Snapchat users.

Topics

- AWS Glue support for Snapchat Ads
- Policies containing the API operations for creating and using connections
- Configuring Snapchat Ads
- Configuring Snapchat Ads connections
- Reading from Snapchat Ads entities
- Snapchat Ads connection options
- Creating a Snapchat Ad account and configuring the client app

- Creating an app in your Snapchat Ads account

AWS Glue support for Snapchat Ads

AWS Glue supports Snapchat Ads as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Snapchat Ads.

Supported as a target?

No.

Supported Snapchat Ads API versions

v1.

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

You can also use the following managed IAM policies to allow access:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, Amazon CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Snapchat Ads

Before you can use AWS Glue to transfer from Snapchat Ads, you must meet these requirements:

Minimum requirements

- You have a Snapchat Ads account. For more information on creating an account, see see Creating a Snapchat Ad account and configuring the client app .
- You have created an OAuth2 app in your Snapchat Ads account. This integration provides the credentials that AWS Glue uses to access your data securely when it makes authenticated calls to your account. For more information, see Creating an app in your Snapchat Ads account .

If you meet these requirements, you’re ready to connect AWS Glue to your Snapchat Ads account. In Snapchat Ads, a connected app is a framework that authorizes external applications, like AWS Glue, to access your Snapchat Ads data.

Configuring Snapchat Ads connections

Snapchat Ads supports only the `AUTHORIZATION_CODE` grant type.

This grant type is considered “three-legged” OAuth as it relies on redirecting users to the third party authorization server to authenticate the user. It is used when creating connections via the AWS Glue Console. The user creating a connection may by default rely on a AWS Glue owned connected app (AWS Glue managed client application) where they do not need to provide any OAuth related information except for their Snapchat Ads instance URL. The AWS Glue Console will redirect the user to Snapchat Ads where the user must login and allow AWS Glue the requested permissions to access their Snapchat Ads instance.

Users may still opt to create their own connected app in Snapchat Ads and provide their own client id and client secret when creating connections through the AWS Glue Console. In this scenario, they will still be redirected to Snapchat Ads to login and authorize AWS Glue to access their resources.

This grant type results in a refresh token and access token. The access token expires after 1 hour of creation. A new access token can be fetched using the refresh token.

For more information on creating a connected app for Authorization Code OAuth flow, see Ads API .

To configure a Snapchat Ads connection:

1. In AWS Secrets Manager, create a secret with the following details. It is required to create a secret for each connection in AWS Glue.
   a. For customer managed connected app – Secret should contain the connected app Consumer Secret with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as key.
2. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Connection type**, select Snapchat Ads.
   b. Provide the Snapchat Ads environment.
   c. Select the IAM role which AWS Glue can assume and has permissions for following actions:

      JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

  d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.

  e. Select the network options if you want to use your network.

3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Snapchat Ads entities

Prerequisites

- A Snapchat Ads Object you would like to read from. Refer the supported entities table below to check the available entities.

Supported entities

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Organization	No	No	No	Yes	No
Ad Account	No	No	No	Yes	No
Creative	No	No	No	Yes	No
Media	No	No	No	Yes	No
Campaign	Yes	No	No	Yes	No
Ad Under Ad Account	Yes	No	No	Yes	No
Ad Under Campaign	No	No	No	Yes	No
Ad Squad	Yes	No	No	Yes	No

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Segment	No	No	No	Yes	No

Example

```
snapchatads_read = glueContext.create_dynamic_frame.from_options(
    connection_type="snapchatAds",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "organization",
        "API_VERSION": "v1"
    }
)
```

Snapchat Ads entity and field details

Snapchat Ads dynamically loads available fields under selected entity. Depending on the data type of the field, it supports following filter operators.

Field Data Type	Supported Filter Operators
Boolean	=

Partitioning queries

- Field-based partitioning: Not supported.
- Record-based partitioning: Not supported.

Snapchat Ads connection options

The following are connection options for Snapchat Ads:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of Snapchat Ads entity. Example: `campaign` .

- `API_VERSION`(String) - (Required) Used for Read. Snapchat Ads Rest API version you want to use. The value will be v1, as Snapchat Ads currently supports only version v1.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Comma separated list of columns you want to select for the selected entity.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.

Creating a Snapchat Ad account and configuring the client app

Topics

- Sign up for Snapchat Ads
- Steps to create a Snapchat Ad account

Sign up for Snapchat Ads

To sign up for Snapchat Ads:

1. Navigate to Snapchat Ads Manager. Choose **Sign Up** next to **New to Snapchat?**.
2. On the **Create Account** screen, follow the prompts to enter your Business Name, Email, Password, etc. Choose **Next**.
3. On the **Create Your Profile** screen, enter values for User Name, Website (Optional), and choose **Create Account**. This will give you an option to add a profile photo and bio on the **Edit Your Profile** screen. Choose **Confirm**.
4. On the **Business Info** screen, fill out the required fields like Country, Currency, Phone Number, GSTIN etc, and complete the account creation process by choosing **Next**.

Steps to create a Snapchat Ad account

To create a Snapchat Ad account:

1. Log in to **Ads Manager**. Then click the navigation in the top corner and select **Ad Accounts**.
2. Choose **+ New Ad Account**. Input your advertiser details:

- Select whether or not you’re an agency buying ads on behalf of an advertiser. If you select ‘Yes’, your ad may be rejected if it uses targeting parameters that could include age, gender, or postal code level targeting. Minimum age targeting may be applied to up to 21 years of age.
- Select whether or not your ad account will run housing, credit, or employment ads. If you select ‘Yes’, your ad may be rejected if it uses targeting parameters that could include age, gender, or postal code level targeting. Minimum age targeting may be applied to up to 21 years of age.
- Select whether you will use the ad account for political ads. If you're running a political ad, input the sponsoring political organization or advocacy group who is paying for the ad. If you do not accurately input the political organization your ads may be rejected. You will also need to fill out the mandatory linked 'Political Ad Review Form' before submitting ads.

3. Choose **Account Details** and fill out your ad account info:

Field	Description
Name	Your ad account name.
Account Type	Select your Account Type, if not auto-popu lated.
Billing Type	If you'd like to continually have access to your funding source, choose 'Revolving.' This allows you to replenish your line of credit and it is often chosen to run ads in Ads Manager. If you'd prefer to run a onetime campaign with a spend cap, you can choose **Insertion Order**. If you create an account with an Insertion order, but want to set a spend cap later, you can leave the spend cap at $0. This way the account will have a spend limit of $0, but you can apply one later.
Advertising Organization	The organization buying ads.

Field	Description
Billing Center	Choose the billing center that you would like to receive invoices. You can use the one automatically created when you first claimed your business or add another Billing Center by following the steps in the Business Help Center .
Currency	Choose your currency.
Spend Cap	If you chose ‘Insertion Order’ as your billing type, input an ad account budget.
Time Zone	Select your time zone.

4. Choose **Create Account**. Your ad account will be created, and you can find it in the Ad Accounts portion of Ads Manager. To begin launching ads, you’ll want to input a payment method. You can also add members to your ad account.
5. Select whether you’d like to use an existing payment or create a new one. Then, choose **Save Payment Method**.
6. Select any members you’ve invited to your business to add to the ad account. For more information about the roles and permissions that can be assigned, see Roles and Permissions Overview. Members added will then be able to log in to Ads Manager and access this ad account. When you’re done, save your members.

For more information about ad accounts, see https://businesshelp.snapchat.com/s/article/roles-permissions?language=en_UShttps://businesshelp.snapchat.com/s/article/roles-permissions?language=en_US

Creating an app in your Snapchat Ads account

To activate access to Snapchat’s Marketing API, make sure you have a business account set up. Then follow the steps below.

1. Log in to Ads Manager. Then choose the menu in the top left corner and select **Business Dashboard**, then select **Business Details**.
2. Choose **+OAuth App** .

3. Enter your App Name and Add following URL as Snap Redirect URI `https://<aws-region>.console.aws.amazon.com/gluestudio/oauth`. For example, if using the us-west-1 region, the URL would be `https://us-west-1.console.aws.amazon.com/gluestudio/oauth) and choose` **`Create OAuth`** **App**. Choose **Create OAuth App**.
4. Your app credentials (Client ID and client Secret) will be displayed. Save them as they will be required to create a connection.

Connecting to Snowflake in AWS Glue Studio

> **Note**
>
> You can use AWS Glue for Spark to read from and write to tables in Snowflake in AWS Glue 4.0 and later versions. To configure a Snowflake connection with AWS Glue jobs programatically, see Redshift connections.

AWS Glue provides built-in support for Snowflake. AWS Glue Studio provides a visual interface to connect to Snowflake, author data integration jobs, and run them on the AWS Glue Studio serverless Spark runtime.

AWS Glue Studio creates a unified connection for Snowflake. For more information, see Considerations.

Topics

- Creating a Snowflake connection
- Creating a Snowflake source node
- Creating a Snowflake target node
- Advanced options

Creating a Snowflake connection

> **Note**
>
> Unified connections (connection v2) standardize all connections to use `USERNAME`, `PASSWORD` keys for basic auth credentials. You can still create a v1 connection via API with secrets containing `sfUser`, `sfPassword`.

When adding a **Data source - Snowflake** node in AWS Glue Studio, you can choose an existing AWS Glue Snowflake connection or create a new connection. You must choose a `SNOWFLAKE` type connection and not a `JDBC` type connection configured to connect to Snowflake. Follow the following procedure to create a AWS Glue Snowflake connection:

To create a Snowflake connection

1. In Snowflake, generate a user, *`snowflakeUser`* and password, *`snowflakePassword`*.
2. Determine which Snowflake warehouse this user will interact with, *`snowflakeWarehouse`*. Either set it as the `DEFAULT_WAREHOUSE` for *`snowflakeUser`* in Snowflake or remember it for the next step.
3. In AWS Secrets Manager, create a secret using your Snowflake credentials. To create a secret in Secrets Manager, follow the tutorial available in [Create an AWS Secrets Manager secret](#) in the AWS Secrets Manager documentation. After creating the secret, keep the Secret name, *`secretName`* for the next step.

   - When selecting **Key/value pairs**, create a pair for *`snowflakeUser`* with the key `sfUser`.
   - When selecting **Key/value pairs**, create a pair for *`snowflakePassword`* with the key `sfPassword`.
   - When selecting **Key/value pairs**, create a pair for *`snowflakeWarehouse`* with the key `sfWarehouse`. This is not needed if a default is set in Snowflake.
4. In the AWS Glue Data Catalog, create a connection by following the steps in [Adding an AWS Glue connection](#). After creating the connection, keep the connection name, *`connectionName`*, for the next step.

   - When selecting a **Connection type**, select Snowflake.
   - When selecting **Snowflake URL**, provide the hostname of your Snowflake instance. The URL will use a hostname in the form *`account_identifier`*`.snowflakecomputing.com`.

- When selecting an **AWS Secret**, provide *`secretName`*.

Creating a Snowflake source node

Permissions needed

AWS Glue Studio jobs using Snowflake data sources require additional permissions. For more information on how to add permissions to ETL jobs, see Review IAM permissions needed for ETL jobs.

`SNOWFLAKE` AWS Glue connections use an AWS Secrets Manager secret to provide credential information. Your job and data preview roles in AWS Glue Studio must have permission to read this secret.

Adding a Snowflake data source

Prerequisites:

- An AWS Secrets Manager secret for your Snowflake credentials
- A Snowflake type AWS Glue Data Catalog connection

To add a Data Source – Snowflake node:

1. Choose the connection for your Snowflake data source. This assumes that the connection already exists and you can select from existing connections. If you need to create a connection, choose **Create Snowflake connection**. For more information, see Overview of using connectors and connections .

   Once you have chosen a connection, you can view the connection properties by clicking **View properties**. Information about the connection are visible, including URL, security groups, subnet, availability zone, description, and created (UTC) and last updated (UTC) timestamps.
2. Choose a Snowflake source option:
   - **Choose a single table** – this is the table that contains the data you want to access from a single Snowflake table.
   - **Enter custom query** – allows you to access a dataset from multiple Snowflake tables based on your custom query.
3. If you chose a single table, enter the name of a Snowflake schema.

Or, choose **Enter custom query**. Choose this option to access a custom dataset from multiple Snowflake tables. When you choose this option, enter the Snowflake query.

4. In **Performance and security** options (optional),

    - **Enable query pushdown** – choose if you want to offload work to the Snowflake instance.

5. In **Custom Snowflake properties** (optional), enter parameters and values as needed.

Creating a Snowflake target node

Permissions needed

AWS Glue Studio jobs using Snowflake data sources require additional permissions. For more information on how to add permissions to ETL jobs, see Review IAM permissions needed for ETL jobs.

`SNOWFLAKE` AWS Glue connections use an AWS Secrets Manager secret to provide credential information. Your job and data preview roles in AWS Glue Studio must have permission to read this secret.

Adding a Snowflake data target

To create a Snowflake target node:

1. Choose an existing Snowflake table as the target, or enter a new table name.
2. When you use the **Data target - Snowflake** target node, you can choose from the following options:

    - **APPEND** – If a table already exists, dump all the new data into the table as an insert. If the table doesn't exist, create it and then insert all new data.
    - **MERGE** – AWS Glue will update or append data to your target table based on the conditions you specify.

      Choose options:

      - **Choose keys and simple actions** – choose the columns to be used as matching keys between the source data and your target data set.

        Specify the following options when matched:

        - Update record in your target data set with data from source.

    - Delete record in your target data set.

    Specify the following options when not matched:

    - Insert source data as a new row into your target data set.
    - Do nothing.
  - **Enter custom MERGE statement** – You can then choose **Validate Merge statement** to verify that the statement is valid or invalid.
- **TRUNCATE** – If a table already exists, truncate the table data by first clearing the contents of the target table. If truncate is successful, then insert all data. If the table doesn't exist, create the table and insert all data. If truncate is not successful, the operation will fail.
- **DROP** – If a table already exists, delete the table metadata and data. If deletion is successful, then insert all data. If the table doesn't exist, create the table and insert all data. If drop is not successful, the operation will fail.

Advanced options

See Snowflake connections in the AWS Glue developer guide.

Connecting to Stripe in AWS Glue Studio

Stripe is an online payment processing and credit card processing platform for businesses. The Stripe platform lets businesses accept online payments, create subscription (recurring billing) for their e-commerce, and sets up back account to receive pay outs. Stripe also supports multi-party payments in which it lets businesses to setup their marketplace and allow to collect payment and then pay out to sellers or service providers through “Connected” account.

Topics

- AWS Glue support for Stripe
- Policies containing the API operations for creating and using connections
- Configuring Stripe
- Configuring Stripe connections
- Reading from Stripe entities
- Stripe connection options
- Limitations
- Creating a new Stripe account and configuring the client app

AWS Glue support for Stripe

AWS Glue supports Stripe as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Stripe.

Supported as a target?

No.

Supported Slack API versions

v1.

Policies containing the API operations for creating and using connections

The following sample policy describes the required IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

You can also use the following managed IAM policies to allow access:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, Amazon CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Stripe

Before you can use AWS Glue to transfer data from Stripe, you must meet these requirements:

Minimum requirements

- You must have a Stripe account with email and password. For more information, see Creating a new Stripe account and configuring the client app.
- Your Stripe account is enabled for API access. All use of the Stripe API is available at no additional cost.

If you meet these requirements, you're ready to connect AWS Glue to your Stripe account.

Configuring Stripe connections

Stripe supports custom authentication. For more information on generating the required API keys for custom authorization, see STRIPE REST API Documentation.

To configure a Stripe connection:

1. In AWS Secrets Manager, create a secret with the following details. It is required to create a secret for each connection in AWS Glue.
   a. For customer managed connected app – Secret should contain the connected app Consumer Secret with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as key.
2. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Connection type**, select Stripe.

   b. Select the IAM role which AWS Glue can assume and has permissions for following actions:

      JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

   c. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.

   d. Select the network options if you want to use your network.

3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Stripe entities

Prerequisites

- A Stripe object you would like to read from.

Supported entities

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Balance	No	No	No	Yes	No

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Balance Transactions	Yes	Yes	No	Yes	Yes
Charges	Yes	Yes	No	Yes	Yes
Disputes	Yes	Yes	No	Yes	Yes
File Links	Yes	Yes	No	Yes	Yes
PaymentIntents	Yes	Yes	No	Yes	Yes
SetupIntents	Yes	Yes	No	Yes	Yes
Payouts	Yes	Yes	No	Yes	Yes
Refunds	Yes	Yes	No	Yes	Yes
Products	Yes	Yes	No	Yes	Yes
Prices	Yes	Yes	No	Yes	Yes
Coupons	Yes	Yes	No	Yes	Yes
Promotion Codes	Yes	Yes	No	Yes	Yes
Tax Codes	No	Yes	No	Yes	No
Tax Rates	Yes	Yes	No	Yes	Yes
Shipping Rates	Yes	Yes	No	Yes	Yes
Sessions	Yes	Yes	No	Yes	Yes
Credit Notes	Yes	Yes	No	Yes	Yes

Entity	Can be Filtered	Supports Limit	Supports Order By	Supports Select *	Supports Partitioning
Customer	Yes	Yes	No	Yes	Yes
Invoices	Yes	Yes	No	Yes	Yes
Invoice Items	Yes	Yes	No	Yes	No
Plans	Yes	Yes	No	Yes	Yes
Quotes	Yes	Yes	No	Yes	No
Subscriptions	Yes	Yes	No	Yes	
Subscription Items	No	Yes	No	Yes	No
Subscription Schedules	Yes	Yes	No	Yes	Yes
Accounts	No	Yes	No	Yes	Yes
Application Fees	Yes	Yes	No	Yes	Yes
Country Specs	No	Yes	No	Yes	No
Transfers	Yes	Yes	No	Yes	Yes
Early Fraud Warnings	Yes	Yes	No	Yes	Yes
Report Types	No	No	No	Yes	No

Example

```
stripe_read = glueContext.create_dynamic_frame.from_options(
    connection_type="stripe",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "coupons",
        "API_VERSION": "v1"
    }
)
```

Stripe entity and field details

Entity	Field	Data Type	Supported Operators
Balance			
	available	List	
	connect_reserved	List	
	pending	List	
	livemode	Boolean	
	object	String	
	instant_available	List	
	issuing	Struct	
Balance Transactions			
	id	String	
	object	String	
	amount	Integer	
	available_on	DateTime	=, >=, <=,<,>
	created	DateTime	=, >=, <=,<,>
	currency	String	

Entity	Field	Data Type	Supported Operators
	description	String	
	exchange_rate	BigDecimal	
	fee	Integer	
	fee_details	List	
	net	Integer	
	reporting_category	String	
	source	String	=
	status	String	
	type	String	=
	cross_border_classification	String	
Charges			
	id	String	
	object	String	
	amount	Integer	=, <, >
	amount_captured	Integer	
	amount_refunded	Integer	
	application	String	
	application_fee	String	
	application_fee_amount	Integer	

Entity	Field	Data Type	Supported Operators
	balance_transaction	String	
	billing_details	Struct	
	calculated_statement_descriptor	String	
	captured	Boolean	
	created	DateTime	=, >=, <=,<,>
	currency	String	
	customer	String	=
	description	String	
	destination	String	
	dispute	String	
	disputed	Boolean	=
	failure_balance_transaction	String	
	failure_code	String	
	failure_message	String	
	fraud_details	Struct	
	invoice	String	
	livemode	Boolean	
	metadata	Struct	
	on_behalf_of	String	

Entity	Field	Data Type	Supported Operators
	order	String	
	outcome	Struct	
	paid	Boolean	
	payment_intent	String	=
	payment_method	String	
	payment_method_details	Struct	
	receipt_email	String	
	receipt_number	String	
	receipt_url	String	
	refunded	Boolean	=
	refunds	Struct	
	review	String	
	shipping	Struct	
	source	Struct	
	source_transfer	String	
	statement_descriptor	String	
	statement_descriptor_suffix	String	
	status	String	
	transfer	String	

Entity	Field	Data Type	Supported Operators
	transfer_data	Struct	
	transfer_group	String	=
Disputes			
	id	String	
	object	String	
	amount	Integer	=, <, >
	balance_transaction	String	
	balance_transactions	List	
	charge	String	=
	created	DateTime	=, >=, <=,<,>
	currency	String	
	evidence	Struct	
	evidence_details	Struct	
	is_charge_refundable	Boolean	
	livemode	Boolean	
	metadata	Struct	
	payment_intent	String	=
	reason	String	=
	status	String	

Entity	Field	Data Type	Supported Operators
	payment_method_details	Struct	
File Links			
	id	String	
	object	String	
	created	DateTime	=, >=, <=,<,>
	expired	Boolean	=
	expires_at	DateTime	
	file	String	=
	livemode	Boolean	
	metadata	Struct	
	url	String	
PaymentIntents			
	id	String	
	object	String	
	amount	Integer	
	amount_capturable	Integer	
	amount_details	Struct	
	amount_received	Integer	
	application	String	

Entity	Field	Data Type	Supported Operators
	application_fee_amount	Integer	
	automatic_payment_methods	Struct	
	canceled_at	DateTime	
	cancellation_reason	String	
	capture_method	String	
	client_secret	String	
	confirmation_method	String	
	created	DateTime	=, >=, <=,<,>
	currency	String	
	customer	String	=
	description	String	
	invoice	String	
	last_payment_error	Struct	
	latest_charge	String	
	livemode	Boolean	
	metadata	Struct	
	next_action	Struct	
	on_behalf_of	String	
	payment_method	String	

Entity	Field	Data Type	Supported Operators
	payment_method_options	Struct	
	payment_method_types	List	
	payment_method_configuration_details	Struct	
	processing	Struct	
	receipt_email	String	
	review	String	
	setup_future_usage	String	
	shipping	Struct	
	source	String	
	statement_descriptor	String	
	statement_descriptor_suffix	String	
	status	String	
	transfer_data	Struct	
	transfer_group	String	
SetupIntents			
	id	String	
	object	String	

Entity	Field	Data Type	Supported Operators
	application	String	
	cancellation_reason	String	
	client_secret	String	
	created	DateTime	=, >=, <=,<,>
	customer	String	=
	description	String	
	flow_directions	List	
	last_setup_error	Struct	
	latest_attempt	String	
	livemode	Boolean	
	mandate	String	
	metadata	Struct	
	next_action	Struct	
	on_behalf_of	String	
	payment_method	String	
	payment_method_options	Struct	
	payment_method_types	List	
	single_use_mandate	String	
	status	String	

Entity	Field	Data Type	Supported Operators
	usage	String	
	automatic_payment_methods	Struct	
Payouts			
	id	String	
	object	String	
	amount	Integer	=, <, >
	arrival_date	DateTime	=, >=, <=,<,>
	automatic	Boolean	
	balance_transaction	String	
	created	DateTime	=, >=, <=,<,>
	currency	String	
	description	String	=
	destination	String	
	failure_balance_transaction	String	
	failure_code	String	
	failure_message	String	
	livemode	Boolean	
	metadata	Struct	
	method	String	

Entity	Field	Data Type	Supported Operators
	original_payout	String	
	reversed_by	String	
	reconciliation_status	String	
	source_type	String	
	statement_descriptor	String	
	status	String	
	type	String	
	application_fee	String	
	application_fee_amount	Integer	
Refunds			
	id	String	
	object	String	
	amount	Integer	
	balance_transaction	String	
	charge	String	=
	created	DateTime	=, >=, <=,<,>
	currency	String	
	metadata	Struct	
	destination_details	Struct	

Entity	Field	Data Type	Supported Operators
	payment_intent	String	=
	reason	String	
	receipt_number	String	
	source_transfer_reversal	String	
	status	String	
	transfer_reversal	String	
Products			
	id	String	
	object	String	
	active	Boolean	=
	attributes	List	
	created	DateTime	=, >=, <=,<,>
	default_price	String	
	description	String	
	images	List	
	livemode	Boolean	
	metadata	Struct	
	name	String	
	package_dimensions	Struct	

Entity	Field	Data Type	Supported Operators
	shippable	Boolean	
	statement_descriptor	String	
	tax_code	String	
	type	String	=
	unit_label	String	
	updated	DateTime	
	url	String	
	features	List	
Prices			
	id	String	
	object	String	
	active	Boolean	=
	billing_scheme	String	
	created	DateTime	=, >=, <=,<,>
	currency	String	=
	custom_unit_amount	Struct	
	livemode	Boolean	
	lookup_key	String	
	metadata	Struct	
	nickname	String	

Entity	Field	Data Type	Supported Operators
	product	String	=
	recurring	Struct	
	tax_behavior	String	
	tiers_mode	String	
	transform_quantity	Struct	
	type	String	=
	unit_amount	Integer	
	unit_amount_decimal	String	
Coupons			
	Id	String	
	object	String	
	amount_off	Integer	
	created	DateTime	=, >=, <=,<,>
	currency	String	=
	duration	String	=
	duration_in_months	Integer	=,<,>
	livemode	Boolean	
	max_redemptions	Integer	=, <, >
	metadata	Struct	
	name	String	

Entity	Field	Data Type	Supported Operators
	percent_off	Double	=
	redeem_by	DateTime	=, >=, <=, <, >
	times_redeemed	Integer	
	valid	Boolean	
Promotion Codes			
	Id	String	
	object	String	
	active	Boolean	=
	code	String	=
	coupon	Struct	
	created	DateTime	=,>=,<=,<,>
	customer	String	
	expires_at	DateTime	
	livemode	Boolean	
	max_redemptions	Integer	
	metadata	Struct	
	restrictions	Struct	
	times_redeemed	Integer	
Tax Codes			
	Id	String	

Entity	Field	Data Type	Supported Operators
	object	String	
	description	String	
	name	String	
Tax Rates			
	Id	String	
	object	String	
	active	Boolean	=
	country	String	
	created	DateTime	=, >=, <=, <, >
	description	String	
	display_name	String	
	inclusive	Boolean	=
	jurisdiction	String	
	jurisdiction_level	String	
	livemode	Boolean	
	metadata	Struct	
	percentage	Double	
	effective_percentage	Double	
	state	String	
	tax_type	String	

Entity	Field	Data Type	Supported Operators
Shipping Rates			
	Id	String	
	object	String	
	active	Boolean	=
	created	DateTime	=, >=, <=, <, >
	delivery_estimate	Struct	
	display_name	String	
	fixed_amount	Struct	
	livemode	Boolean	
	metadata	Struct	
	tax_behavior	String	
	tax_code	String	
	type	String	
Sessions			
	id	String	
	object	String	
	after_expiration	Struct	
	allow_promotion_codes	Boolean	
	amount_subtotal	Integer	

Entity	Field	Data Type	Supported Operators
	amount_total	Integer	
	automatic_tax	Struct	
	billing_address_collection	String	
	cancel_url	String	
	client_reference_id	String	
	consent	Struct	
	consent_collection	Struct	
	created	DateTime	=, >=, <=, <, >
	currency	String	
	custom_text	Struct	
	customer	String	
	customer_creation	String	
	customer_details	Struct	
	customer_email	String	
	expires_at	DateTime	
	invoice	String	
	invoice_creation	Struct	
	livemode	Boolean	
	locale	String	

Entity	Field	Data Type	Supported Operators
	metadata	Struct	
	mode	String	
	payment_intent	String	=
	payment_link	String	
	payment_method_collection	String	
	payment_method_options	Struct	
	payment_method_types	List	
	payment_status	String	
	phone_number_collection	Struct	
	recovered_from	String	
	setup_intent	String	
	shipping_address_collection	Struct	
	shipping_cost	Struct	
	shipping_details	Struct	
	shipping_options	List	
	status	String	
	submit_type	String	

Entity	Field	Data Type	Supported Operators
	subscription	String	
	success_url	String	
	tax_id_collection	Struct	
	total_details	Struct	
	url	String	
	ui_mode	String	
Credit Notes			
	id	String	
	object	String	
	amount	Integer	
	created	DateTime	=, >=, <=, <, >
	currency	String	
	customer	String	=
	customer_balance_transaction	String	
	discount_amount	Integer	
	discount_amounts	List	
	invoice	String	=
	lines	Struct	
	livemode	Boolean	

Entity	Field	Data Type	Supported Operators
	memo	String	
	metadata	Struct	
	number	String	
	out_of_band_amount	Integer	
	pdf	String	
	reason	String	
	refund	String	
	status	String	
	subtotal	Integer	
	subtotal_excluding_tax	Integer	
	tax_amounts	List	
	total	Integer	
	total_excluding_tax	Integer	
	type	String	
	voided_at	DateTime	
	amount_shipping	Integer	
	effective_at	DateTime	
	shipping_cost	Struct	
Customer			

Entity	Field	Data Type	Supported Operators
	id	String	
	object	String	
	address	Struct	
	balance	Integer	
	created	DateTime	
	currency	String	=, >=, <=, <, >
	default_source	String	
	delinquent	Boolean	=
	description	String	
	discount	Struct	
	email	String	=
	invoice_prefix	String	
	invoice_settings	Struct	
	livemode	Boolean	
	metadata	Struct	
	name	String	
	next_invoice_sequence	Integer	
	phone	String	
	preferred_locales	List	

Entity	Field	Data Type	Supported Operators
	shipping	Struct	
	tax_exempt	String	
	test_clock	String	
Invoices			
	id	String	
	object	String	
	account_country	String	
	account_name	String	
	account_tax_ids	List	
	amount_due	Integer	
	amount_paid	Integer	
	amount_remaining	Integer	
	application	String	
	application_fee_amount	Integer	
	attempt_count	Integer	
	attempted	Boolean	=
	auto_advance	Boolean	=
	automatic_tax	Struct	
	billing_reason	String	

Entity	Field	Data Type	Supported Operators
	charge	String	
	collection_method	String	=
	created	DateTime	=, >=, <=, <, >
	currency	String	
	custom_fields	List	
	customer	String	=
	customer_address	Struct	
	customer_email	String	
	customer_name	String	
	customer_phone	String	
	customer_shipping	Struct	
	customer_tax_exempt	String	
	customer_tax_ids	List	
	default_payment_method	String	
	default_source	String	
	default_tax_rates	List	
	description	String	
	discount	Struct	
	discounts	List	

Entity	Field	Data Type	Supported Operators
	due_date	DateTime	=, >=, <=, <, >
	ending_balance	Integer	
	footer	String	
	from_invoice	Struct	
	hosted_invoice_url	String	
	invoice_pdf	String	
	last_finalization_error	Struct	
	latest_revision	String	
	lines	Struct	
	livemode	Boolean	
	metadata	Struct	
	next_payment_attempt	DateTime	
	number	String	
	on_behalf_of	String	
	paid	Boolean	=
	paid_out_of_band	Boolean	
	payment_intent	String	
	payment_settings	Struct	
	period_end	DateTime	=, >=, <=, <, >

Entity	Field	Data Type	Supported Operators
	period_start	DateTime	=, >=, <=, <, >
	post_payment_credit_notes_amount	Integer	
	pre_payment_credit_notes_amount	Integer	
	quote	String	
	receipt_number	String	
	rendering	Struct	
	rendering_options	Struct	
	starting_balance	Integer	
	statement_descriptor	String	
	status	String	=
	status_transitions	Struct	
	subscription	String	
	subscription_details	Struct	
	subtotal	Integer	=, <, >
	subtotal_excluding_tax	Integer	
	tax	Integer	
	test_clock	String	
	total	Integer	=, <, >

Entity	Field	Data Type	Supported Operators
	total_discount_amounts	List	
	total_excluding_tax	Integer	
	total_tax_amounts	List	
	transfer_data	Struct	
	webhooks_delivered_at	DateTime	
	automatically_finalizes_at	DateTime	
	effective_at	DateTime	
	issuer	Struct	
Invoice Items			
	id	String	
	object	String	
	amount	Integer	=, <, >
	currency	String	
	customer	String	=
	date	DateTime	
	description	String	
	discountable	Boolean	
	discounts	List	

Entity	Field	Data Type	Supported Operators
	invoice	String	=
	livemode	Boolean	
	metadata	Struct	
	period	Struct	
	plan	Struct	
	price	Struct	
	proration	Boolean	=
	quantity	Integer	
	subscription	String	
	subscription_item	String	
	tax_rates	List	
	test_clock	String	
	unit_amount	Integer	
	unit_amount_decimal	String	
Plans			
	id	String	
	object	String	
	active	Boolean	=
	aggregate_usage	String	
	amount	Integer	

Entity	Field	Data Type	Supported Operators
	amount_decimal	String	
	billing_scheme	String	
	created	DateTime	=, >=, <=, <, >
	currency	String	=
	interval	String	=
	interval_count	Integer	
	livemode	Boolean	
	metadata	Struct	
	nickname	String	
	product	String	=
	tiers_mode	String	
	transform_usage	Struct	
	trial_period_days	Integer	=, <, >
	usage_type	String	
	meter	String	
Quotes			
	id	String	
	object	String	
	amount_subtotal	Integer	
	amount_total	Integer	

Entity	Field	Data Type	Supported Operators
	application	String	
	application_fee_amount	Integer	
	application_fee_percent	Double	
	automatic_tax	Struct	
	collection_method	String	
	computed	Struct	
	created	DateTime	
	currency	String	
	customer	String	=
	default_tax_rates	List	
	description	String	
	discounts	List	
	expires_at	DateTime	
	footer	String	
	from_quote	Struct	
	header	String	
	invoice	String	
	invoice_settings	Struct	
	livemode	Boolean	

Entity	Field	Data Type	Supported Operators
	metadata	Struct	
	number	String	
	on_behalf_of	String	
	status	String	=
	status_transitions	Struct	
	subscription	String	
	subscription_data	Struct	
	subscription_schedule	String	
	test_clock	String	
	total_details	Struct	
	transfer_data	Struct	
Subscriptions			
	id	String	
	object	String	
	application	String	
	application_fee_percent	Double	
	automatic_tax	Struct	
	billing_cycle_anchor	DateTime	
	billing_thresholds	Struct	

Entity	Field	Data Type	Supported Operators
	cancel_at	DateTime	
	cancel_at_period_end	Boolean	
	canceled_at	DateTime	
	collection_method	String	=
	created	DateTime	=, >=, <=,<,>
	currency	String	
	current_period_end	DateTime	=, >=, <=
	current_period_start	DateTime	=, >=, <=
	customer	String	=
	days_until_due	Integer	
	default_payment_method	String	
	default_source	String	
	default_tax_rates	List	
	description	String	
	discount	Struct	
	ended_at	DateTime	
	items	Struct	
	latest_invoice	String	
	livemode	Boolean	

Entity	Field	Data Type	Supported Operators
	metadata	Struct	
	next_pending_invoice_item_invoice	DateTime	
	pause_collection	Struct	
	payment_settings	Struct	
	pending_invoice_item_interval	Struct	
	pending_setup_intent	String	
	pending_update	Struct	
	plan	Struct	
	quantity	Integer	
	schedule	String	
	start_date	DateTime	
	status	String	=
	test_clock	String	
	transfer_data	Struct	
	trial_end	DateTime	
	trial_start	DateTime	
Subscription Items			
	Id	String	

Entity	Field	Data Type	Supported Operators
	object	String	
	billing_thresholds	Struct	
	created	DateTime	=, >=, <=, <, >
	metadata	Struct	
	plan	Struct	
	price	Struct	
	subscription	String	
	tax_rates	List	
	discounts	List	
Subscription Schedules			
	object	String	
	application	String	
	canceled_at	DateTime	
	completed_at	DateTime	
	created	DateTime	
	current_phase	Struct	
	customer	String	=
	default_settings	Struct	
	end_behavior	String	

Entity	Field	Data Type	Supported Operators
	livemode	Boolean	
	metadata	Struct	
	phases	List	
	released_at	DateTime	
	released_subscription	String	
	renewal_interval	String	
	status	String	
	subscription	String	
	test_clock	String	
Accounts			
	details_submitted	Boolean	
	tos_acceptance	Struct	
	type	String	
	metadata	Struct	
	id	String	
	object	String	
	default_currency	String	
	capabilities	Struct	
	charges_enabled	Boolean	
	settings	Struct	

Entity	Field	Data Type	Supported Operators
	requirements	Struct	
	payouts_enabled	Boolean	
	future_requirements	Struct	
	external_accounts	Struct	
	controller	Struct	
	country	String	
	email	String	
	created	DateTime	=, >=, <=, <, >
	business_profile	Struct	
	business_type	String	
	company	Struct	
Application Fees			
	id	String	
	object	String	
	account	String	
	amount	Integer	=, <, >
	amount_refunded	Integer	=, <, >
	application	String	
	balance_transaction	String	
	charge	String	=

Entity	Field	Data Type	Supported Operators
	created	DateTime	
	currency	String	
	livemode	Boolean	
	originating_transaction	String	
	refunded	Boolean	=
	refunds	Struct	
	fee_source	Struct	
Country Specs			
	id	String	
	object	String	
	default_currency	String	
	supported_bank_account_currencies	Struct	
	supported_payment_currencies	List	
	supported_payment_methods	List	
	supported_transfer_countries	List	
	verification_fields	Struct	
Transfers			

Entity	Field	Data Type	Supported Operators
	id	String	
	object	String	
	amount	Integer	=, <, >
	amount_reversed	Integer	
	balance_transaction	String	
	created	DateTime	=, >=, <=, <, >
	currency	String	=
	description	String	
	destination	String	=
	destination_payment	String	
	livemode	Boolean	
	metadata	Struct	
	reversals	Struct	
	reversed	Boolean	
	source_transaction	String	
	source_type	String	
	transfer_group	String	=
Early Fraud Warnings			
	id	String	
	object	String	

Entity	Field	Data Type	Supported Operators
	actionable	Boolean	
	charge	String	=
	created	DateTime	=, >=, <=, <, >
	fraud_type	String	
	livemode	Boolean	
	payment_intent	String	=
Report Types			
	id	String	
	object	String	
	data_available_end	DateTime	
	data_available_start	DateTime	
	default_columns	List	
	livemode	Boolean	
	name	String	
	updated	DateTime	
	version	Integer	

Partitioning queries

Additional spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, `NUM_PARTITIONS` can be provided if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition query.
- `LOWER_BOUND`: an inclusive lower bound value of the chosen partition field.

  For date, we accept the Spark date format used in Spark SQL queries. Example of valid value: `"2024-07-01T00:00:00.000Z"`.
- `UPPER_BOUND`: an exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: number of partitions.

Entity wise partitioning field support details are captured in below table.

Entity Name	Partitioning Field	Data Type
Balance Transactions	created	DateTime
Charges	created	DateTime
Disputes	created	DateTime
File Links	created	DateTime
PaymentIntents	created	DateTime
SetupIntents	created	DateTime
Payouts	created	DateTime
Refunds	created	DateTime
Products	created	DateTime
Prices	created	DateTime
Coupons	created	DateTime
Promotion Codes	created	DateTime
Tax Rates	created	DateTime
Shipping Rates	created	DateTime

Entity Name	Partitioning Field	Data Type
Sessions	created	DateTime
Credit Notes	created	DateTime
Customer	created	DateTime
Invoices	created	DateTime
Plans	created	DateTime
Subscriptions	created	DateTime
Subscription Schedules	created	DateTime
Accounts	created	DateTime
Application Fees	created	DateTime
Transfers	created	DateTime
Early Fraud Warnings	created	DateTime

Example

```
stripe_read = glueContext.create_dynamic_frame.from_options(
    connection_type="stripe",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "coupons",
        "API_VERSION": "v1",
        "PARTITION_FIELD": "created"
        "LOWER_BOUND": "2024-05-01T20:55:02.000Z"
        "UPPER_BOUND": "2024-07-11T20:55:02.000Z"
        "NUM_PARTITIONS": "10"
    }
)
```

Stripe connection options

The following are connection options for Stripe:

- `ENTITY_NAME`(String) - (Required) Used for Read/Write. The name of your Object in Stripe.
- `API_VERSION`(String) - (Required) Used for Read/Write. Stripe Rest API version you want to use. Example: v1.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.

Limitations

The following are limitations for the Stripe connector:

- Only Field Based Partitioning supported by connector.
- Record Based Partitioning not supported by connector, no provision to retrieve the total count of records.
- Primary key datatype is String, so Id Based Partitioning does not support by connector.

Creating a new Stripe account and configuring the client app

Creating a Stripe account

1. Choose on the link **https://dashboard.stripe.com/register**.
2. Enter your Email, Full name, Password & choose **Create Account**.
3. After login with account, verify the account by choosing **Open Gmail**.

4. Verify the account by clicking on the verification link received on email.
5. After clicking on verify email address, it will redirect to another page
6. After clicking on **Activate payments** to activate the account, it will redirect to Activate payments (**https://dashboard.stripe.com/welcome**) page and make sure to fill all your valid details and after that choose **Continue** button.

Creating a Slack developer app

1. Log in to Stripe.
2. Choose **Developers** as displayed on extreme top of the picture below.
3. Choose **API keys** under Developers.
4. Choose **Reveal test key** to get the API keys.

Connecting to Teradata Vantage in AWS Glue Studio

AWS Glue provides built-in support for Teradata Vantage. AWS Glue Studio provides a visual interface to connect to Teradata, author data integration jobs, and run them on the AWS Glue Studio serverless Spark runtime.

AWS Glue Studio creates a unified connection for Teradata Vantage. For more information, see Considerations.

Topics

- Creating a Teradata Vantage connection
- Creating a Teradata source node
- Creating a Teradata target node
- Advanced options

Creating a Teradata Vantage connection

To connect to Teradata Vantage from AWS Glue, you will need to create and store your Teradata credentials in an AWS Secrets Manager secret, then associate that secret with a AWS Glue Teradata connection.

Prerequisites:

- If you are accessing your Teradata environment through Amazon VPC, configure Amazon VPC to allow your AWS Glue job to communicate with the Teradata environment. We discourage accessing the Teradata environment over the public internet.

  In Amazon VPC, identify or create a **VPC**, **Subnet** and **Security group** that AWS Glue will use while executing the job. Additionally, you need to ensure Amazon VPC is configured to permit network traffic between your Teradata instance and this location. Your job will need to establish a TCP connection with your Teradata client port. For more information about Teradata ports, see the Teradata documentation.

  Based on your network layout, secure VPC connectivity may require changes in Amazon VPC and other networking services. For more information about AWS connectivity, consult AWS Connectivity Options in the Teradata documentation.

To configure a AWS Glue Teradata connection:

1. In your Teradata configuration, identify or create a user and password AWS Glue will connect with, *`teradataUser`* and *`teradataPassword`*. For more information, consult Vantage Security Overview in the Teradata documentation.
2. In AWS Secrets Manager, create a secret using your Teradata credentials. To create a secret in Secrets Manager, follow the tutorial available in Create an AWS Secrets Manager secret in the AWS Secrets Manager documentation. After creating the secret, keep the Secret name, *`secretName`* for the next step.
   - When selecting **Key/value pairs**, create a pair for the key `user` with the value *`teradataUsername`*.
   - When selecting **Key/value pairs**, create a pair for the key `password` with the value *`teradataPassword`*.
3. In the AWS Glue console, create a connection by following the steps in the section called “Adding an AWS Glue connection”. After creating the connection, keep the connection name, *`connectionName`*, for the next step.
   - When selecting a **Connection type**, select Teradata.
   - When providing **JDBC URL**, provide the URL for your instance. You can also hardcode certain comma separated connection parameters in your JDBC URL. The URL must conform to the following format: `jdbc:teradata://teradataHostname/ParameterName=ParameterValue,ParameterName`

Supported URL parameters include:

  - `DATABASE`– name of database on host to access by default.
  - `DBS_PORT`– the database port, used when running on a nonstandard port.
- When selecting a **Credential type**, select **AWS Secrets Manager**, then set **AWS Secret** to *`secretName`*.

4. In the following situations, you may require additional configuration:

   - For Teradata instances hosted on AWS in an Amazon VPC
     - You will need to provide Amazon VPC connection information to the AWS Glue connection that defines your Teradata security credentials. When creating or updating your connection, set **VPC**, **Subnet** and **Security groups** in **Network options**.

Creating a Teradata source node

Prerequisites needed

- An AWS Glue Teradata Vantage connection, configured with an AWS Secrets Manager secret, as described in the previous section, the section called "Creating a Teradata Vantage connection".
- Appropriate permissions on your job to read the secret used by the connection.
- A Teradata table you would like to read from, *`tableName`*, or query *`targetQuery`*.

Adding a Teradata data source

To add a Data source – Teradata node:

1. Choose the connection for your Teradata data source. Since you have created it, it should be available in the dropdown. If you need to create a connection, choose **Create a new connection**. For more information see the previous section, the section called "Creating a Teradata Vantage connection".

   Once you have chosen a connection, you can view the connection properties by clicking **View properties**.
2. Choose a **Teradata Source** option:

   - **Choose a single table** – access all data from a single table.
   - **Enter custom query** – access a dataset from multiple tables based on your custom query.

3. If you chose a single table, enter *tableName*.

   If you chose **Enter custom query**, enter a SQL SELECT query.
4. In **Custom Teradata properties**, enter parameters and values as needed.

Creating a Teradata target node

Prerequisites needed

- A AWS Glue Teradata Vantage connection, configured with an AWS Secrets Manager secret, as described in the previous section, the section called "Creating a Teradata Vantage connection".
- Appropriate permissions on your job to read the secret used by the connection.
- A Teradata table you would like to write to, *tableName*.

Adding a Teradata data target

To add a Data target – Teradata node:

1. Choose the connection for your Teradata data source. Since you have created it, it should be available in the dropdown. If you need to create a connection, choose **Create Teradata connection**. For more information, see Overview of using connectors and connections .

   Once you have chosen a connection, you can view the connection properties by clicking **View properties**.
2. Configure **Table name** by providing *tableName*.
3. In **Custom Teradata properties**, enter parameters and values as needed.

Advanced options

You can provide advanced options when creating a Teradata node. These options are the same as those available when programming AWS Glue for Spark scripts.

See the section called "Teradata Vantage connections".

Connecting to Twilio

Twilio provides programmable communication tools for making and receiving phone calls, sending, and receiving text messages, and performing other communication functions using its web service

APIs. Twilio's APIs power its platform for communications. Behind these APIs is a software layer connecting and optimizing communications networks around the world to allow your users to call and message anyone, globally. As a Twilio user, you can connect AWS Glue to your Twilio account. Then, you can use Twilio as a data source in your ETL jobs. Run these jobs to transfer data between Twilio and AWS services or other supported applications.

Topics

- AWS Glue support for Twilio
- Policies containing the API operations for creating and using connections
- Configuring Twilio
- Configuring Twilio connections
- Reading from Twilio entities
- Twilio connection options
- Limitations and notes for Twilio connector

AWS Glue support for Twilio

AWS Glue supports Twilio as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Twilio.

Supported as a target?

No.

Supported Twilio API versions

The following Twilio API versions are supported:

- v1
- 2010-04-01

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Twilio

Before you can use AWS Glue to transfer data from Twilio, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- You have a Twilio account with username and password.
- Your Twilio account is enabled for API access.

If you meet these requirements, you’re ready to connect AWS Glue to your Twilio account. For typical connections, you don't need do anything else in Twilio.

Configuring Twilio connections

Twilio supports username and password for Basic Authentication. Basic Authentication is a simple authentication method where clients provide credentials directly to access protected resources. AWS Glue is able to use the username (Account SID) and password (Auth Token) to authenticate Twilio APIs.

For public Twilio documentation for Basic Authentication flow, see Basic Authentication | Twilio.

To configure a Twilio connection:

1. In AWS Secrets Manager, create a secret with the following details:
    - For Basic Authentication: the Secret should contain the connected app Consumer Secret with the **Account SID** (Username) and **Auth Token** (Password).

    > **Note**
    >
    > You must create a secret for your connections in AWS Glue.

2. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
    1. When selecting a **Connection type**, select Twilio.
    2. Provide the `Edge_Location` of the Twilio instance you want to connect to.
    3. Select the AWS IAM role which AWS Glue can assume and has permissions for following actions:

        JSON

        ```
        {
          "Version": "2012-10-17",
          "Statement": [
            {
              "Effect": "Allow",
        ```

```
        "Action": [
          "secretsmanager:DescribeSecret",
          "secretsmanager:GetSecretValue",
          "secretsmanager:PutSecretValue",
          "ec2:CreateNetworkInterface",
          "ec2:DescribeNetworkInterfaces",
          "ec2:DeleteNetworkInterface"
        ],
        "Resource": "*"
      }
    ]
  }
```

   d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.

   e. Select the network options if you want to use your network.

3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Twilio entities

Prerequisite

A Twilio object you would like to read from. You will need the object name such as `SMS-Message` or `SMS-CountryPricing`.

Supported entities for source:

Entity	Interface	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
SMS-Message	REST	Yes	Yes	No	Yes	Yes
SMS-CountryPricing	REST	No	No	No	Yes	No
Voice-Call	REST	Yes	Yes	No	Yes	No

Entity	Interface	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Voice-Application	REST	Yes	Yes	No	Yes	No
Voice-OutgoingCallerID	REST	Yes	Yes	No	Yes	No
Voice-Queue	REST	Yes	Yes	No	Yes	No
Conversations-Conversation	REST	Yes	Yes	No	Yes	No
Conversations-User	REST	No	Yes	No	Yes	No
Conversations-Role	REST	No	Yes	No	Yes	No
Conversations-Configuration	REST	No	No	No	Yes	No
Conversations-AddressConfiguration	REST	Yes	Yes	No	Yes	No
Conversations-WebhookConfiguration	REST	No	No	No	Yes	No

Entity	Interface	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Conversations-ParticipantConversation	REST	No	No	No	Yes	No
Conversations-Credential	REST	No	Yes	No	Yes	No
Conversations-ConversationService	REST	No	Yes	No	Yes	No

Example:

```
twilio_read = glueContext.create_dynamic_frame.from_options(
    connection_type="twilio",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "sms-message",
        "API_VERSION": "2010-04-01",
        "Edge_Location": "sydney.us1"
    }
```

Twilio entity and field details:

Entity	Field	Data type	Supported operators
SMS-Message	account_sid	String	N/A
	api_version	String	N/A
	body	String	N/A

Entity	Field	Data type	Supported operators
	date_created	Datetime	N/A
	date_sent	Datetime	>=, <=, =
	date_updated	Datetime	N/A
	direction	String	N/A
	error_code	Integer	N/A
	error_message	String	N/A
	from	Integer	=
	messaging_service_sid	String	N/A
	num_media	String	N/A
	num_segments	String	N/A
	price	String	N/A
	price_unit	Struct	N/A
	sid	Integer	N/A
	status	String	N/A
	subresource_uris	Map	N/A
	to	Integer	=
	uri	Datetime	N/A
SMS-CountryPricing	country	String	N/A
	iso_country	String	N/A
	url	String	N/A

Entity	Field	Data type	Supported operators
	outbound_sms_prices	List	N/A
	inbound_sms_prices	List	N/A
	price_unit	String	N/A
Voice-Call	account_sid	String	N/A
	annotation	String	N/A
	answered_by	String	N/A
	api_version	String	N/A
	caller_name	String	N/A
	date_created	Datetime	N/A
	date_updated	Datetime	N/A
	direction	String	N/A
	duration	String	N/A
	end_time	Datetime	>=, <=, =
	forwarded_from	String	N/A
	from	String	=
	from_formatted	String	N/A
	group_sid	String	N/A
	parent_call_sid	String	N/A
	phone_number_sid	String	N/A
	price	String	N/A

Entity	Field	Data type	Supported operators
	price_unit	String	N/A
	sid	String	N/A
	start_time	Datetime	>=, <=, =
	status	String	=
	subresource_uris	String	N/A
	to	String	=
	to_formatted	String	N/A
	trunk_sid	String	N/A
	uri	String	N/A
	queue_time	String	N/A
Voice-Application	account_sid	String	N/A
	api_version	String	N/A
	date_created	Datetime	N/A
	date_updated	Datetime	N/A
	friendly_name	String	=
	message_status_callback	String	N/A
	sid	String	N/A
	sms_fallback_method	String	N/A
	sms_fallback_url	String	N/A
	sms_method	String	N/A

Entity	Field	Data type	Supported operators
	sms_status_callback	String	N/A
	sms_url	String	N/A
	status_callback	String	N/A
	status_callback_method	String	N/A
	uri	String	N/A
	voice_caller_id_lookup	Boolean	N/A
	voice_fallback_method	String	N/A
	voice_fallback_url	String	N/A
	voice_method	String	N/A
	voice_url	String	N/A
public_application_connect_enabled	Boolean	N/A	
Voice-OutgoingCallerID	sid	String	N/A
	date_created	Datetime	N/A
	date_updated	Datetime	N/A
	account_sid	String	N/A
	friendly_name	String	=
	phone_number	String	=
	uri	String	N/A

Entity	Field	Data type	Supported operators
Voice-Queue	date_created	Datetime	N/A
	date_updated	Datetime	N/A
	current_size	Integer	N/A
	friendly_name	String	N/A
	uri	String	N/A
	account_sid	String	N/A
	average_wait_time	Integer	N/A
	sid	String	N/A
	max_size	Integer	N/A
Conversations-Conversation	account_sid	String	N/A
	chat_service_sid	String	N/A
	messaging_service_sid	String	N/A
	sid	String	N/A
	friendly_name	String	N/A
	unique_name	String	N/A
	attributes	String	N/A
	state	String	=
	date_created	Datetime	N/A
	date_updated	Datetime	N/A
	timers	Struct	N/A

Entity	Field	Data type	Supported operators
	url	String	N/A
	links	Struct	N/A
	bindings	Struct	N/A
	start_date	Datetime	=
	end_date	Datetime	=
	Timers.DateInactive	String	N/A
	Timers.DateClosed	String	N/A
Conversations-User	sid	String	N/A
	account_sid	String	N/A
	chat_service_sid	String	N/A
	role_sid	String	N/A
	identity	String	N/A
	friendly_name	String	N/A
	attributes	String	N/A
	is_online	Boolean	N/A
	is_notifiable	Boolean	N/A
	date_created	Datetime	N/A
	date_updated	Datetime	N/A
	url	String	N/A
	links	Struct	N/A

Entity	Field	Data type	Supported operators
Conversations-Role	sid	String	N/A
	account_sid	String	N/A
	chat_service_sid	String	N/A
	friendly_name	String	N/A
	type	String	N/A
	permissions	String	N/A
	date_created	Datetime	N/A
	date_updated	Datetime	N/A
	url	String	N/A
Conversations-Configuration	account_sid	Long	N/A
	default_chat_service_sid	String	N/A
	default_messaging_service_sid	String	N/A
	default_inactive_timer	String	N/A
	default_closed_timer	String	N/A
	url	String	N/A
	links	Map	N/A
Conversations-AddressConfiguration	sid	String	N/A
	account_sid	String	N/A
	type	String	N/A

Entity	Field	Data type	Supported operators
	address	String	N/A
	friendly_name	String	N/A
	auto_creation	Struct	N/A
	date_created	Datetime	N/A
	date_updated	Datetime	N/A
	url	String	N/A
	address_country	String	N/A
	AutoCreation.Enabled	Boolean	N/A
	AutoCreation.Type	String	N/A
	AutoCreation.ConversationServiceSid	String	N/A
	AutoCreation.WebhookUrl	String	N/A
	AutoCreation.WebhookMethod	String	N/A
	AutoCreation.WebhookFilters	List	N/A
	AutoCreation.StudioFlowSid	String	N/A
	AutoCreation.StudioRetryCount	Integer	N/A
Conversations-WebhookConfiguration	account_sid	String	N/A
	method	String	N/A

Entity	Field	Data type	Supported operators
	filters	List	N/A
	pre_webhook_url	String	N/A
	post_webhook_url	String	N/A
	target	String	N/A
	url	String	N/A
Converations-ParticipantConversation	account_sid	String	N/A
	chat_service_sid	String	N/A
	participant_sid	String	N/A
	participant_user_sid	String	N/A
	participant_identity	String	N/A
	participant_messaging_binding	Struct	N/A
	Conversation_sid	String	N/A
	conversation_unique_name	String	N/A
	conversation_friendly_name	String	N/A
	conversation_attributes	String	N/A
	conversation_date_created	Datetime	N/A
	conversation_date_updated	Datetime	N/A

Entity	Field	Data type	Supported operators
	conversation_created_by	String	N/A
	conversation_state	String	N/A
	conversation_timers	Struct	N/A
	links	Map	N/A
	address	String	=
	identity	String	=
Conversation-Credentials	sid	String	N/A
	account_sid	String	N/A
	friendly_name	String	N/A
	type	String	N/A
	sandbox	String	N/A
	date_created	Datetime	N/A
	dated_updated	Datetime	N/A
	url	String	N/A
	certificate	String	N/A
	private_key	String	N/A
	api_key	String	N/A
	secret	String	N/A
Conversations-ConversationService	sid	String	N/A
	account_sid	String	N/A

Entity	Field	Data type	Supported operators
	friendly_name	String	N/A
	date_created	Datetime	N/A
	date_updated	Datetime	N/A
	url	String	N/A
	links	Map	N/A

Partitioning queries

Fields supporting partitioning:

In Twilio, the DateTime datatype fields support field-based partitioning.

You can provide the additional Spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, and `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition the query.
- `LOWER_BOUND`: an **inclusive** lower bound value of the chosen partition field.

  For the Datetime field, we accept the Spark timestamp format used in Spark SQL queries.

  Examples of valid value:

  ```
  "2024-05-01T20:55:02.000Z"
  ```
- `UPPER_BOUND`: an **exclusive** upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: the number of partitions.

Example:

```
twilio_read = glueContext.create_dynamic_frame.from_options(
    connection_type="twilio",
```

```
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "sms-message",
        "API_VERSION": "2010-04-01",
        "PARTITION_FIELD": "date_sent"
        "LOWER_BOUND": "2024-05-01T20:55:02.000Z"
        "UPPER_BOUND": "2024-06-01T20:55:02.000Z"
        "NUM_PARTITIONS": "10"
    }
```

Twilio connection options

The following are connection options for Twilio:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your object in Twilio.
- `EDGE_LOCATION`(String) - (Required) A valid Twilio edge location.
- `API_VERSION`(String) - (Required) Used for Read. Twilio Rest API version you want to use. Twilio supports two API versions: ‘v1’ and ‘2010-04-01’.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for Read.
- `INSTANCE_URL`(String) - (Required) Used for Read. A valid Twilio instance URL.

Limitations and notes for Twilio connector

The following are limitations or notes for the Twilio connector:

- Record-based partitioning is not supported, as there is no provision to retrieve the total count of records from Twilio.

- The fields `date_sent`, `start_time`, and `end_time` are of the Datetime datatype, but when filtering, they only support date values (time components are not considered).
- Filtering the "from" or "to" fields works only if the values do not include any prefix (for example, a protocol or label). If a prefix is present, filtering for the respective field does not work. For example if you pass "to": "whatsapp:+14xxxxxxxxxx" as a filter then Twilio won't return a response. You need to pass it as "to": "+14xxxxxxxx", then it will return records if they exist.
- The "identity" field filter is mandatory when querying the `conversation-participant-conversation` entity.

Connecting to Vertica in AWS Glue Studio

AWS Glue provides built-in support for Vertica. AWS Glue Studio provides a visual interface to connect to Vertica, author data integration jobs, and run them on the AWS Glue Studio serverless Spark runtime.

AWS Glue Studio creates a unified connection for Vertica. For more information, see Considerations.

Topics

- Creating a Vertica connection
- Creating a Vertica source node
- Creating a Vertica target node
- Advanced options

Creating a Vertica connection

Prerequisites:

- An Amazon S3 bucket or folder to use for temporary storage when reading from and writing to the database, referred to by *`tempS3Path`*.

> **Note**
>
> When using Vertica in AWS Glue job data previews, temporary files may not be automatically removed from *`tempS3Path`*. To ensure the removal of temporary files, directly end the data preview session by choosing **End session** in the **Data preview** pane.

If you cannot guarantee the data preview session is ended directly, consider setting Amazon S3 Lifecycle configuration to remove old data. We recommend removing data older than 49 hours, based on maximum job runtime plus a margin. For more information about configuring Amazon S3 Lifecycle, see Managing your storage lifecycle in the Amazon S3 documentation.

- An IAM policy with appropriate permissions to your Amazon S3 path you can associate with your AWS Glue job role.
- If your Vertica instance is in an Amazon VPC, configure Amazon VPC to allow your AWS Glue job to communicate with the Vertica instance without traffic traversing the public internet.

  In Amazon VPC, identify or create a **VPC**, **Subnet** and **Security group** that AWS Glue will use while executing the job. Additionally, you need to ensure Amazon VPC is configured to permit network traffic between your Vertica instance and this location. Your job will need to establish a TCP connection with your Vertica client port, (default 5433). Based on your network layout, this may require changes to security group rules, Network ACLs, NAT Gateways and Peering connections.

To configure a connection to Vertica:

1. In AWS Secrets Manager, create a secret using your Vertica credentials, *`verticaUsername`* and *`verticaPassword`*. To create a secret in Secrets Manager, follow the tutorial available in Create an AWS Secrets Manager secret in the AWS Secrets Manager documentation. After creating the secret, keep the Secret name, *`secretName`* for the next step.
   - When selecting **Key/value pairs**, create a pair for the key `user` with the value *`verticaUsername`*.
   - When selecting **Key/value pairs**, create a pair for the key `password` with the value *`verticaPassword`*.
2. In the AWS Glue console, create a connection by following the steps in the section called "Adding an AWS Glue connection". After creating the connection, keep the connection name, *`connectionName`*, for the next step.
   - When selecting a **Connection type**, select Vertica.
   - When selecting **Vertica Host**, provide the hostname of your Vertica installation.
   - When selecting **Vertica Port**, the port your Vertica installation is available through.

  - When selecting an **AWS Secret**, provide *`secretName`*.

3. In the following situations, you may require additional configuration:

   - For Vertica instances hosted on AWS in an Amazon VPC
     - Provide Amazon VPC connection information to the AWS Glue connection that defines your Vertica security credentials. When creating or updating your connection, set **VPC**, **Subnet** and **Security groups** in **Network options**.

You will need to perform the following steps before running your AWS Glue job:

- Grant the IAM role associated with your AWS Glue job permissions to *`tempS3Path`*.
- Grant the IAM role associated with your AWS Glue job permission to read *`secretName`*.

Creating a Vertica source node

Prerequisites needed

- A Vertica type AWS Glue Data Catalog connection, *`connectionName`* and a temporary Amazon S3 location, *`tempS3Path`*, as described in the previous section, the section called “Creating a Vertica connection”.
- A Vertica table you would like to read from, *`tableName`*, or query *`targetQuery`*.

Adding a Vertica data source

To add a Data source – Vertica node:

1. Choose the connection for your Vertica data source. Since you have created it, it should be available in the dropdown. If you need to create a connection, choose **Create Vertica connection**. For more information see the previous section, the section called “Creating a Vertica connection”.

   Once you have chosen a connection, you can view the connection properties by clicking **View properties**.
2. Choose the **Database** containing your table.
3. Choose the **Staging area in Amazon S3**, enter an S3A URI to *`tempS3Path`*.
4. Choose the **Vertica Source**.

  - **Choose a single table** – access all data from a single table.
  - **Enter custom query** – access a dataset from multiple tables based on your custom query.

5. If you chose a single table, enter *tableName* and optionally select a **Schema**.

   If you chose **Enter custom query**, enter a SQL SELECT query and optionally select a **Schema**.
6. In **Custom Vertica properties**, enter parameters and values as needed.

Creating a Vertica target node

Prerequisites needed

- A Vertica type AWS Glue Data Catalog connection, *connectionName* and a temporary Amazon S3 location, *tempS3Path*, as described in the previous section, the section called "Creating a Vertica connection".

Adding a Vertica data target

To add a Data target – Vertica node:

1. Choose the connection for your Vertica data source. Since you have created it, it should be available in the dropdown. If you need to create a connection, choose **Create Vertica connection**. For more information see the previous section, the section called "Creating a Vertica connection".

   Once you have chosen a connection, you can view the connection properties by clicking **View properties**.
2. Choose the **Database** containing your table.
3. Choose the **Staging area in Amazon S3**, enter an S3A URI to *tempS3Path*.
4. Enter *tableName* and optionally select a **Schema**.
5. In **Custom Vertica properties**, enter parameters and values as needed.

Advanced options

You can provide advanced options when creating a Vertica node. These options are the same as those available when programming AWS Glue for Spark scripts.

See the section called “Vertica connections”.

Connecting to WooCommerce

WooCommerce is an open-source flexible software solution built for WordPress-based websites. It's commonly used to create online e-commerce shops. With this software solution, anyone can turn their regular website into a fully-functioning online store.

Topics

- AWS Glue support for WooCommerce
- Policies containing the API operations for creating and using connections
- Configuring WooCommerce
- Configuring WooCommerce connections
- Reading from WooCommerce entities
- WooCommerce connection options

AWS Glue support for WooCommerce

AWS Glue supports WooCommerce as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from WooCommerce.

Supported as a target?

No.

Supported WooCommerce API versions

The following WooCommerce API versions are supported:

- v3

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring WooCommerce

Before you can use AWS Glue to transfer data from WooCommerce, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- You have a WooCommerce account with a `consumerKey` and a `consumerSecret`.
- Your WooCommerce account has API access with a valid license.

If you meet these requirements, you’re ready to connect AWS Glue to your WooCommerce account. For typical connections, you don't need do anything else in WooCommerce.

Configuring WooCommerce connections

WooCommerce supports custom authentication. For public WooCommerce documentation on generating the required API keys for custom authorization, see Authentication – WooCommerce REST API Documentation.

To configure a WooCommerce connection:

1. In AWS Secrets Manager, create a secret with the following details:
   - For a customer managed connected app, the Secret should contain the connected app Consumer Secret with `consumerKey` and `consumerSecret` as keys. Note: you must create a secret per connection in AWS Glue.

1. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Connection type**, select WooCommerce.
   b. Provide the `INSTANCE_URL` of the WooCommerce instance you want to connect to.
   c. Select the AWS IAM role which AWS Glue can assume and has permissions for following actions:

      JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
          "ec2:DeleteNetworkInterface"
        ],
        "Resource": "*"
      }
    ]
  }
```

d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.

e. Select the network options if you want to use your network.

2. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from WooCommerce entities

Prerequisite

A WooCommerce object you would like to read from. You will need the object name such as coupon, order, product, etc.

Supported entities for source:

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Coupon	Yes	Yes	Yes	Yes	Yes
Coupon Total	No	No	No	Yes	No
Customers Total	No	No	No	Yes	No
Order	Yes	Yes	Yes	Yes	Yes
Orders Total	No	No	No	Yes	No
Payment Gateway	No	No	No	Yes	No
Product	Yes	Yes	Yes	Yes	Yes

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Product attribute	Yes	Yes	Yes	Yes	Yes
Product category	Yes	Yes	Yes	Yes	Yes
Product review	Yes	Yes	Yes	Yes	Yes
Product shipping class	Yes	Yes	Yes	Yes	Yes
Product tag	Yes	Yes	Yes	Yes	Yes
Product variation	Yes	Yes	Yes	Yes	Yes
Products Total	No	No	No	Yes	No
Report (List)	No	No	No	Yes	No
Reviews Total	No	No	No	Yes	No
Sales Report	Yes	No	No	Yes	No
Shipping Method	No	No	No	Yes	No
Shipping Zone	No	No	No	Yes	No
Shipping Zone Location	No	No	No	Yes	No

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Shipping Zone Method	No	No	No	Yes	No
Tax Rate	Yes	Yes	Yes	Yes	Yes
Tax Class	No	No	No	Yes	No
Top Sellers Report	Yes	No	No	Yes	No

Example:

```
woocommerce_read = glueContext.create_dynamic_frame.from_options(
    connection_type="glue.spark.woocommerce",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "coupon",
        "API_VERSION": "v3",
        "INSTANCE_URL": "instanceUrl"
    }
```

WooCommerce entity and field details:

Entity	Field	Data type	Supported operators
coupon	id	Integer	N/A
	code	String	EQUAL_TO
	amount	String	N/A
	status	String	N/A
	date_created	DateTime	N/A
	date_created_gmt	DateTime	N/A

Entity	Field	Data type	Supported operators
	date_modified	DateTime	N/A
	date_modified_gmt	DateTime	N/A
	discount_type	String	N/A
	description	String	N/A
	date_expires	String	N/A
	date_expires_gmt	String	N/A
	usage_count	Integer	N/A
	individual_use	Boolean	N/A
	product_ids	List	N/A
	excluded_product_ids	List	N/A
	usage_limit	Integer	N/A
	usage_limit_per_user	Integer	N/A
	limit_usage_to_x_items	Integer	N/A
	free_shipping	Boolean	N/A
	product_categories	List	N/A
	excluded_product_categories	List	N/A
	exclude_sale_items	Boolean	N/A
	minimum_amount	String	N/A
	maximum_amount	String	N/A

Entity	Field	Data type	Supported operators
	email_restrictions	List	N/A
	used_by	List	N/A
	meta_data	List	N/A
	context	String	EQUAL_TO
	search	String	EQUAL_TO
	after	DateTime	EQUAL_TO
	before	DateTime	EQUAL_TO
	order	String	EQUAL_TO
	orderby	String	EQUAL_TO
	modified_after	DateTime	EQUAL_TO
	modified_before	DateTime	EQUAL_TO
	dates_are_gmt	Boolean	EQUAL_TO
coupon-total	slug	String	N/A
	name	String	N/A
	total	Integer	N/A
customer-total	slug	String	N/A
	name	String	N/A
	total	Integer	N/A
order	id	Integer	N/A
	parent_id	Integer	N/A

Entity	Field	Data type	Supported operators
	number	String	N/A
	order_key	String	N/A
	created_via	String	N/A
	status	String	N/A
	currency	String	N/A
	version	String	N/A
	date_created	DateTime	N/A
	date_modified	DateTime	N/A
	discount_total	String	N/A
	discount_tax	String	N/A
	shipping_total	String	N/A
	shipping_tax	String	N/A
	cart_tax	String	N/A
	total	String	N/A
	total_tax	String	N/A
	prices_include_tax	Boolean	N/A
	customer_id	Integer	N/A
	customer_ip_address	String	N/A
	customer_user_agent	String	N/A
	customer_note	String	N/A

Entity	Field	Data type	Supported operators
	billing	Struct	N/A
	shipping	Struct	N/A
	payment_method	String	N/A
	payment_method_title	String	N/A
	transaction_id	String	N/A
	date_paid	DateTime	N/A
	date_completed	DateTime	N/A
	cart_hash	String	N/A
	meta_data	List	N/A
	line_items	List	N/A
	tax_lines	List	N/A
	shipping_lines	List	N/A
	fee_lines	List	N/A
	coupon_lines	List	N/A
	refunds	List	N/A
	payment_url	String	N/A
	is_editable	Boolean	N/A
	needs_payment	Boolean	N/A
	needs_processing	Boolean	N/A
	date_created_gmt	DateTime	N/A

Entity	Field	Data type	Supported operators
	date_modified_gmt	DateTime	N/A
	date_completed_gmt	DateTime	N/A
	date_paid_gmt	DateTime	N/A
	currency_symbol	String	N/A
	set_paid	Boolean	N/A
	context	String	EQUAL_TO
	search	String	EQUAL_TO
	after	DateTime	EQUAL_TO
	before	DateTime	EQUAL_TO
	order	String	EQUAL_TO
	orderby	String	EQUAL_TO
	customer	Integer	EQUAL_TO
	product	Integer	EQUAL_TO
	dp	Integer	EQUAL_TO
	modified_before	DateTime	EQUAL_TO
	modified_after	DateTime	EQUAL_TO
	dates_are_gmt	Boolean	EQUAL_TO
order-total	slug	String	N/A
	name	String	N/A
	total	Integer	N/A

Entity	Field	Data type	Supported operators
payment-gateway	title	String	N/A
	description	String	N/A
	order	String	N/A
	enabled	Boolean	N/A
	method_title	String	N/A
	method_description	String	N/A
	method_supports	List	N/A
	settings	String	N/A
	needs_setup	Boolean	N/A
	post_install_scripts	List	N/A
	settings_url	String	N/A
	connection_url	String	N/A
	setup_help_text	String	N/A
	required_settings_keys	List	N/A
product	id	Integer	N/A
	name	String	N/A
	type	String	EQUAL_TO
	permalink	String	N/A
	date_created	DateTime	N/A
	date_created_gmt	DateTime	N/A

Entity	Field	Data type	Supported operators
	date_modified	DateTime	N/A
	date_modified_gmt	DateTime	N/A
	catalog_visibility	String	N/A
	description	String	N/A
	short_description	String	N/A
	price	String	N/A
	regular_price	String	N/A
	sale_price	String	N/A
	date_on_sale_from	DateTime	N/A
	date_on_sale_from_gmt	DateTime	N/A
	date_on_sale_to	DateTime	N/A
	date_on_sale_to_gmt	DateTime	N/A
	price_html	String	N/A
	purchasable	Boolean	N/A
	total_sales	Integer	N/A
	virtual	Boolean	N/A
	downloadable	Boolean	N/A
	downloads	List	N/A
	download_limit	Integer	N/A
	download_expiry	Integer	N/A

Entity	Field	Data type	Supported operators
	external_url	String	N/A
	button_text	String	N/A
	tax_status	String	N/A
	manage_stock	Boolean	N/A
	stock_quantity	Integer	N/A
	backorders	String	N/A
	backorders_allowed	Boolean	N/A
	backordered	Boolean	N/A
	sold_individually	Boolean	N/A
	weight	String	N/A
	dimensions	Struct	N/A
	shipping_required	Boolean	N/A
	shipping_taxable	Boolean	N/A
	shipping_class_id	Integer	N/A
	reviews_allowed	Boolean	N/A
	average_rating	String	N/A
	rating_count	Integer	N/A
	related_ids	List	N/A
	upsell_ids	List	N/A
	cross_sell_ids	List	N/A

Entity	Field	Data type	Supported operators
	parent_id	Integer	N/A
	purchase_note	String	N/A
	categories	List	N/A
	tags	List	N/A
	images	List	N/A
	attributes	List	N/A
	default_attributes	List	N/A
	variations	List	N/A
	grouped_products	List	N/A
	menu_order	Integer	N/A
	meta_data	List	N/A
	low_stock_amount	Integer	N/A
	jetpack_publicize_connections	List	N/A
	jetpack-related-posts	List	N/A
	jetpack_likes_enabled	Boolean	N/A
	jetpack_sharing_enabled	Boolean	N/A
	context	String	EQUAL_TO
	search	String	EQUAL_TO
	after	DateTime	EQUAL_TO

Entity	Field	Data type	Supported operators
	before	DateTime	EQUAL_TO
	order	String	EQUAL_TO
	orderby	String	EQUAL_TO
	slug	String	EQUAL_TO
	status	String	EQUAL_TO
	sku	String	EQUAL_TO
	featured	Boolean	EQUAL_TO
	tag	String	EQUAL_TO
	shipping_class	String	EQUAL_TO
	tax_class	String	EQUAL_TO
	on_sale	Boolean	EQUAL_TO
	stock_status	String	EQUAL_TO
	has_options	Boolean	N/A
	modified_after	DateTime	EQUAL_TO
	modified_before	DateTime	EQUAL_TO
	dates_are_gmt	Boolean	EQUAL_TO
	category	String	EQUAL_TO
	attribute	String	EQUAL_TO
	min_price	String	EQUAL_TO
	max_price	String	EQUAL_TO

Entity	Field	Data type	Supported operators
product-attribute	id	Integer	N/A
	name	String	N/A
	slug	String	N/A
	type	String	N/A
	order_by	String	N/A
	has_archives	Boolean	N/A
	context	String	EQUAL_TO
product-attribute-term	id	Integer	N/A
	name	String	N/A
	slug	String	N/A
	description	String	N/A
	menu_order	Integer	N/A
	count	Integer	N/A
	context	String	EQUAL_TO
	search	String	EQUAL_TO
	order	String	EQUAL_TO
	orderby	String	EQUAL_TO
	hide_empty	Boolean	EQUAL_TO
	parent	Integer	EQUAL_TO
	product	Integer	EQUAL_TO

Entity	Field	Data type	Supported operators
product-category	id	Integer	N/A
	name	String	N/A
	slug	String	EQUAL_TO
	description	String	N/A
	display	String	N/A
	image	Struct	N/A
	menu_order	Integer	N/A
	count	Integer	N/A
	context	String	EQUAL_TO
	search	String	EQUAL_TO
	order	String	EQUAL_TO
	orderby	String	EQUAL_TO
	hide_empty	Boolean	EQUAL_TO
	parent	Integer	EQUAL_TO
	product	Integer	EQUAL_TO
product-review	id	Integer	N/A
	date_created	DateTime	N/A
	date_created_gmt	DateTime	N/A
	product_id	Integer	N/A
	product_name	String	N/A

Entity	Field	Data type	Supported operators
	product_permalink	String	N/A
	review	String	N/A
	rating	Integer	N/A
	verified	Boolean	N/A
	reviewer	String	N/A
	reviewer_email	String	N/A
	reviewer_avatar_urls	Struct	N/A
	context	String	EQUAL_TO
	search	String	EQUAL_TO
	after	DateTime	EQUAL_TO
	before	DateTime	EQUAL_TO
	order	String	EQUAL_TO
	orderby	String	EQUAL_TO
	status	String	EQUAL_TO
product-shipping-class	id	Integer	N/A
	name	String	N/A
	slug	String	EQUAL_TO
	description	String	N/A
	count	Integer	N/A
	context	String	EQUAL_TO

Entity	Field	Data type	Supported operators
	search	String	EQUAL_TO
	order	String	EQUAL_TO
	orderby	String	EQUAL_TO
	hide_empty	String	EQUAL_TO
	product	Integer	EQUAL_TO
product-tag	id	Integer	N/A
	name	String	N/A
	slug	String	EQUAL_TO
	description	String	N/A
	count	Integer	N/A
	context	String	EQUAL_TO
	search	String	EQUAL_TO
	order	String	EQUAL_TO
	orderby	String	EQUAL_TO
	hide_empty	Boolean	EQUAL_TO
	product	Integer	EQUAL_TO
product-total	slug	String	N/A
	name	String	N/A
	total	Integer	N/A
product-variation	id	Integer	N/A

Entity	Field	Data type	Supported operators
	date_created	DateTime	N/A
	date_created_gmt	DateTime	N/A
	date_modified	DateTime	N/A
	date_modified_gmt	DateTime	N/A
	description	String	N/A
	permalink	String	N/A
	price	String	N/A
	regular_price	String	N/A
	sale_price	String	N/A
	date_on_sale_from	DateTime	N/A
	date_on_sale_from_gmt	DateTime	N/A
	date_on_sale_to	DateTime	N/A
	date_on_sale_to_gmt	DateTime	N/A
	purchasable	Boolean	N/A
	virtual	Boolean	N/A
	downloadable	Boolean	N/A
	downloads	List	N/A
	download_limit	Integer	N/A
	download_expiry	Integer	N/A
	tax_status	String	N/A

Entity	Field	Data type	Supported operators
	manage_stock	Boolean	N/A
	stock_quantity	Integer	N/A
	backorders	String	N/A
	backorders_allowed	Boolean	N/A
	backordered	Boolean	N/A
	low_stock_amount	Integer	N/A
	weight	String	N/A
	dimensions	Struct	N/A
	shipping_class	String	N/A
	shipping_class_id	Integer	N/A
	image	Struct	N/A
	attributes	List	N/A
	menu_order	Integer	N/A
	meta_data	List	N/A
	context	String	EQUAL_TO
	search	String	EQUAL_TO
	after	DateTime	EQUAL_TO
	before	DateTime	EQUAL_TO
	order	String	EQUAL_TO
	orderby	String	EQUAL_TO

Entity	Field	Data type	Supported operators
	slug	String	EQUAL_TO
	status	String	EQUAL_TO
	sku	String	EQUAL_TO
	tax_class	String	EQUAL_TO
	on_sale	Boolean	EQUAL_TO
	min_price	String	EQUAL_TO
	max_price	String	EQUAL_TO
	stock_status	String	EQUAL_TO
report	slug	String	N/A
	description	String	N/A
review-total	slug	String	N/A
	name	String	N/A
	total	Integer	N/A
sales-report	total_sales	String	N/A
	net_sales	String	N/A
	average_sales	String	N/A
	total_orders	Integer	N/A
	total_items	Integer	N/A
	total_tax	String	N/A
	total_shipping	String	N/A

Entity	Field	Data type	Supported operators
	total_refunds	Integer	N/A
	total_discount	String	N/A
	totals_grouped_by	String	N/A
	totals	Struct	N/A
	total_customers	Integer	N/A
	context	String	EQUAL_TO
	period	String	EQUAL_TO
	date_min	Date	EQUAL_TO
	date_max	Date	EQUAL_TO
shipping-method	id	String	N/A
	title	String	N/A
	description	String	N/A
shipping-zone	id	Integer	EQUAL_TO
	name	String	N/A
	order	Integer	N/A
shipping-zone-location	code	String	N/A
	type	String	N/A
shipping-zone-method	instance_id	Integer	N/A
	id	Integer	EQUAL_TO
	title	String	N/A

Entity	Field	Data type	Supported operators
	order	Integer	N/A
	enabled	Boolean	N/A
	method_id	String	N/A
	method_title	String	N/A
	method_description	String	N/A
	settings	Struct	N/A
tax-class	slug	String	N/A
	name	String	N/A
tax-rate	id	Integer	N/A
	country	String	N/A
	state	String	N/A
	postcode	String	N/A
	city	String	N/A
	postcodes	List	N/A
	cities	List	N/A
	rate	String	N/A
	name	String	N/A
	priority	Integer	N/A
	compound	Boolean	N/A
	shipping	Boolean	N/A

Entity	Field	Data type	Supported operators
	context	String	EQUAL_TO
	order	String	EQUAL_TO
	orderby	String	EQUAL_TO
	class	String	EQUAL_TO
top-seller-report	name	String	N/A
	product_id	Integer	N/A
	quantity	Integer	N/A
	context	String	EQUAL_TO
	period	String	EQUAL_TO
	date_min	Date	EQUAL_TO
	date_max	Date	EQUAL_TO

> **Note**
>
> Struct and List data types are converted to String data type, and DateTime data type is converted to Timestamp in the response of the connectors.

Partitioning queries

Record-based partitioning:

You can provide the additional Spark option `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

In record-based partitioning, the total number of records present is queried from the WooCommerce API, and divided by a `NUM_PARTITIONS` number provided. The resulting number of records are then concurrently fetched by each sub-query.

- `NUM_PARTITIONS`: the number of partitions.

The following entities support record-based partitioning:

- coupon
- order
- product
- product-attribute
- product-attribute-term
- product-category
- product-review
- product-shipping-class
- product-tag
- product-variation
- tax-rate

Example:

```
woocommerce_read = glueContext.create_dynamic_frame.from_options(
    connection_type="glue.spark.woocommerce",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "coupon",
        "API_VERSION": "v3",
        "INSTANCE_URL": "instanceUrl"
        "NUM_PARTITIONS": "10"
    }
```

Record-based partitioning:

The original query is splitinto `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently:

- `NUM_PARTITIONS`: the number of partitions.

Example:

```
WooCommerce_read = glueContext.create_dynamic_frame.from_options(
    connection_type="WooCommerce",
    connection_options={
        "connectionName": "connectionName",
        "REALMID": "1234567890123456789",
        "ENTITY_NAME": "Bill",
        "API_VERSION": "v3",
        "NUM_PARTITIONS": "10"
    }
```

WooCommerce connection options

The following are connection options for WooCommerce:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your object in WooCommerce.
- `API_VERSION`(String) - (Required) Used for Read. WooCommerce Rest API version you want to use.
- `REALM_ID`(String) - An ID that identifies an individual WooCommerce Online company where you send requests.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `INSTANCE_URL`(String) - (Required) A valid WooCommerce instance URL with the format: https://<instance>.wpcomstaging.com
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.

Connecting to Zendesk

Zendesk is a cloud-based help desk management solution offering customizable tools to build a customer service portal, knowledge base and online communities.

Topics

- AWS Glue support for Zendesk
- Policies containing the API operations for creating and using connections
- Configuring Zendesk
- Configuring Zendesk connections
- Reading from Zendesk entities
- Zendesk connection options
- Limitations

AWS Glue support for Zendesk

AWS Glue supports Zendesk as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Zendesk.

Supported as a target?

No.

Supported Zendesk API versions

The following Zendesk API versions are supported

- v2

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you dont want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Zendesk

Before you can use AWS Glue to transfer data from Zendesk, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- You have a Zendesk account. For more information, see Creating a Zendesk account.
- Your Zendesk account is enabled for API access.
- Your Zendesk account allows you to install connected apps.

If you meet these requirements, you’re ready to connect AWS Glue to your Zendesk account.

Creating a Zendesk account

To create a Zendesk account:

1. Go to https://www.zendesk.com/in/register/
2. Enter the details such as your work email, first name, last name, phone number, job title, company name, number of employees in company, password and preferred Language. Then choose **Complete trial Signup**.
3. Once your account is created, complete the verification link you received to verify your email address.
4. Once the work email address is verified, you are redirected to your Zendesk account. Choose the **Buy Zendesk option** for your preferred plan. Note: for the Zendesk connector it is recommended to purchase the Suite Enterprise plan.

Creating a client app and OAuth 2.0 credentials

To create a client app and OAuth 2.0 credentials:

1. Log into your Zendesk account where you want the OAuth 2.0 app to be created https://www.zendesk.com/in/login/
2. Click the gear icon. Choose the **Go to admin center** link to open the admin center page.
3. Choose **Apps and integrations** in the left sidebar, then select **APIs** > **Zendesk API**.
4. On the Zendesk API page, choose the **OAuth Clients** tab.
5. Choose **Add Oauth Client** on the right side.
6. Complete the following fields to create a client:
    a. Client Name - Enter a name for your app. This is the name that users will see when asked to grant access to your application, and when they check the list of third-party apps that have access to their Zendesk.
    b. Description - Optional. A short description of your app that users will see when asked to grant access to it.
    c. Company - Optional. The company name that users will see when asked to grant access to your application. The information can help them understand who they're granting access to.
    d. Logo - Optional. This is the logo that users will see when asked to grant access to your application. The image can be JPG, GIF, or PNG. For best results, upload a square image. It will be resized for the authorization page.

   e. Unique Identifier - The field is auto-populated with a reformatted version of the name you entered for your app. You can change it if you want.
   f. Redirect URLs - Enter the URL or URLs that Zendesk should use to send the user's decision to grant access to your application.

      For example: https://us-east-1.console.aws.amazon.com/gluestudio/oauth

7. Click **Save**.
8. After the page refreshes, a new pre-populated **Secret** field appears on the lower side. This is the "client_secret" value specified in the OAuth2 spec. Copy the Secret value to your clipboard and save it somewhere safe. Note: The characters may extend past the width of the text box, so make sure to select everything before copying.
9. Click **Save**.

Configuring Zendesk connections

The Zendesk connector supports the Authorization Code grant type.

- This grant type is considered "three-legged" OAuth as it relies on redirecting users to a third-party authorization server to authenticate the user. It is used when creating connections via the AWS Glue console. The user creating a connection may by default rely on an AWS Glue-owned connected app (AWS Glue-managed client application) where they do not need to provide any OAuth related information except for their Zendesk instance URL. The AWS Glue console will redirect the user to Zendesk where the user must login and allow AWS Glue the requested permissions to access their Zendesk instance.
- You may still opt to create your own connected app in Zendesk and provide your own client ID and client secret when creating connections through the AWS Glue console. In this scenario, you will still be redirected to Zendesk to login and authorize AWS Glue to access your resources.
- This grant type results in an access token. The access token never expires.

For public Zendesk documentation on creating a connected app for the Authorization Code OAuth flow, see OAuth Tokens for Grant Types.

To configure a Zendesk connection:

1. In AWS Secrets Manager, create a secret with the following details:

   a. For the AuthorizationCode grant type: for a customer managed connected app the secret should contain the connected app Consumer Secret with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as key.

   b. Note: You must create a secret per connection in AWS Glue.

2. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:

   a. When selecting a **Connection type**, select Zendesk.

   b. Provide the INSTANCE_URL of the Zendesk you want to connect to.

   c. Provide the Zendesk environment.

   d. Select the AWS IAM role which AWS Glue can assume and has permissions for following actions:

      JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

   e. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.

   f. Select the network options if you want to use your network.

3. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Zendesk entities

Prerequisite

A Zendesk Object you would like to read from. You will need the object name such as ticket or user or article, as mentioned in the following table.

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Ticket	Y	Y	Y	Y	N
User	Y	Y	Y	Y	N
Organization	Y	Y	Y	Y	N
Article	Y	Y	N	Y	N
Ticket Event	Y	Y	N	Y	N
Ticket Metric Event	Y	Y	N	Y	N
Ticket Comment	Y	Y	Y	Y	N
Ticket Field	Y	Y	N	Y	N
Ticket Metric	Y	Y	N	Y	N
Ticket Activity	Y	Y	N	Y	N
Ticket Skip	N	Y	N	Y	N
Group	Y	Y	Y	Y	N
Group Membership	N	Y	Y	Y	N

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Satisfaction Rating	Y	Y	N	Y	N
View	Y	Y	Y	Y	N
Trigger	Y	Y	Y	Y	N
Trigger Category	N	Y	Y	Y	N
Macro	Y	Y	Y	Y	N
Automation	N	Y	Y	Y	N

Example:

```
Zendesk_read = glueContext.create_dynamic_frame.from_options(
    connection_type="Zendesk",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "Account",
        "API_VERSION": "v2"
    }
```

Zendesk entities and field details:

Entity	Field	Data type	Supported operators	Comments
articles	url	String		
	id	Long		
	author_id	Long		
	body	String		

Entity	Field	Data type	Supported operators	Comments
	comments_disabled	Boolean		
	draft	Boolean		
	edited_at	DateTime		
	html_url	String		
	label_names	List		
	locale	String	EQUAL_TO	
	outdated	Boolean		
	outdated_locales	List		
	permission_group_id	Long		
	position	Integer		
	promoted	Boolean		
	section_id	Long		
	source_locale	String		
	name	String		
	title	String		
	user_segment_id	Long		
	content_tags_id	List		
	vote_count	Integer		

Entity	Field	Data type	Supported operators	Comments
	vote_sum	Integer		
	created_at	DateTime		
	updated_at	DateTime	EQUAL_TO	
	label_name	String	EQUAL_TO	
group	url	String		
	id	Long		
	is_public	Boolean		
	name	String		
	description	String		
	default	Boolean		
	deleted	Boolean		
	created_at	DateTime		
	updated_at	DateTime		
	exclude_deleted	Boolean	EQUAL_TO	
automation	url	String		
	id	Long		
	title	String		
	active	Boolean		
	created_at	DateTime		
	updated_at	DateTime		

Entity	Field	Data type	Supported operators	Comments
	default	Boolean		
	actions	List		
	positions	Integer		
	conditions	Struct		
	raw_title	String		
group-membership	url	String		
	id	Long		
	user_id	Long		
	group_id	Long		
	default	Boolean		
	created_at	DateTime		
	updated_at	DateTime		
macro	url	String		
	id	Long		
	title	String		
	active	Boolean	EQUAL_TO	
	created_at	DateTime		
	updated_at	DateTime		
	default	Boolean		
	actions	List		

Entity	Field	Data type	Supported operators	Comments
	position	Integer		
	description	String		
	raw_title	String		
	restriction	Struct		
	access	String	EQUAL_TO	
	category	Integer	EQUAL_TO	
	group_id	Long	EQUAL_TO	
	only_viewable	Boolean	EQUAL_TO	
organizations	url	String		
	id	Long		
	external_id	String		
	name	String		
	domain_names	List		
	details	String		
	notes	String		
	group_id	Long		
	shared_tickets	Boolean		
	shared_comments	Boolean		
	tags	List		

Entity	Field	Data type	Supported operators	Comments
	organization_fields	Struct		
	created_at	DateTime		
	updated_at	DateTime	EQUAL_TO	
	DML_STATUS	String		A user-defined field used to track the created, updated and deleted status of the record.
satisfaction-rating	url	String		
	id	Long		
	assignee_id	Long		
	comment	String		
	group_id	Long		
	reason	String		
	reason_code	Integer		
	reason_id	Long		
	requester_id	Long		
	score	String	EQUAL_TO	
	ticket_id	Integer		
	created_at	DateTime		

Entity	Field	Data type	Supported operators	Comments
	updated_at	DateTime	EQUAL_TO	
	start_time	DateTime	EQUAL_TO	
	end_time	DateTime	EQUAL_TO	
	DML_STATUS	String		A user-defined field used to track the created, updated and deleted status of the record.
ticket-activity	actor	Struct		
	actor_id	Long		
	created_at	DateTime		
	id	Long		
	object	Struct		
	target	Struct		
	title	String		
	updated_at	DateTime		
	url	String		
	user	Struct		
	user_id	Long		
	verb	String		

Entity	Field	Data type	Supported operators	Comments
	since	DateTime	EQUAL_TO	
ticket-comment	id	Long		
	type	String		
	author_id	Long		
	body	String		
	html_body	String		
	plain_body	String		
	public	Boolean		
	attachments	List		
	audit_id	Long		
	via	Struct		
	created_at	DateTime		
	metadata	Struct		
	ticket_id	Integer	EQUAL_TO	
	include_inline_images	Boolean	EQUAL_TO	
ticket-events	id	Long		
	ticket_id	Long		
	timestamp	Long		
	created_at	DateTime		

Entity	Field	Data type	Supported operators	Comments
	updater_id	Long		
	child_events	List		
	via	String		
	system	Struct		
	event_type	String		
	comment_present	Boolean		
	comment_public	Boolean		
	via_reference_id	Long		
	created_at	DateTime	EQUAL_TO	
	DML_STATUS	String		A user-defined field used to track the created, updated and deleted status of the record.
ticket-field	url	String		
	id	Long		
	type	String		
	title	String		
	raw_title	String		
	description	String		

Entity	Field	Data type	Supported operators	Comments
	raw_description	String		
	position	Integer		
	active	Boolean		
	required	Boolean		
	collapsed_for_agents	Boolean		
	regexp_for_validation	String		
	title_in_portal	String		
	raw_title_in_portal	String		
	visible_in_portal	Boolean		
	editable_on_portal	Boolean		
	required_in_portal	Boolean		
	tag	String		
	created_at	DateTime		
	updated_at	DateTime		
	removable	Boolean		
	agent_description	String		

Entity	Field	Data type	Supported operators	Comments
	custom_field_options	List		
	custom_statuses	List		
	relationship_filter	Struct		
	relationship_target_type	String		
	sub_type_id	Integer		
	system_field_options	List		
	locale	String	EQUAL_TO	
ticket-metric-events	id	Long		
	time	DateTime	EQUAL_TO	
	ticket_id	Integer		
	metric	String		
	instance_id	Integer		
	type	String		
	DML_STATUS	String	EQUAL_TO	A user-defined field used to track the created, updated and deleted status of the record.

Entity	Field	Data type	Supported operators	Comments
ticket-metric	url	String		
	id	Long		
	ticket_id	Integer		
	created_at	DateTime		
	updated_at	DateTime		
	group_stations	Integer		
	assignee_stations	Integer		
	reopens	Integer		
	replies	Integer		
	assignee_updated_at	DateTime		
	requester_updated_at	DateTime		
	initially_assigned_at	DateTime		
	assigned_at	DateTime		
	solved_at	DateTime		
	last_comment_added_at	DateTime		
	reply_time_in_minutes	Struct		

Entity	Field	Data type	Supported operators	Comments
	first_resolution_time_in_minutes	Struct		
	full_resolution_time_in_minutes	Struct		
	agent_wait_time_in_minutes	Struct		
	requester_wait_time_in_minutes	Struct		
	on_hold_time_in_seconds	Struct		
	reply_time_in_seconds	Struct		
	custom_status_updated_at	DateTime		
ticket-skip	created_at	DateTime		
	id	Long		
	reason	String		
	ticket	Struct		
	ticket_id	Integer		
	updated_at	DateTime		

Entity	Field	Data type	Supported operators	Comments
	user_id	Long		
tickets	url	String		
	id	Long		
	external_id	String	EQUAL_TO	
	type	String		
	subject	String		
	raw_subject	String		
	description	String		
	priority	String		
	status	String		
	recipient	String		
	requester	Struct		
	requester_id	Long		
	submitter_id	Long		
	assignee_id	Long		
	organization_id	Long		
	group_id	Long		
	collaborator_ids	List		
	emails_cc_ids	List		
	follower_ids	List		

Entity	Field	Data type	Supported operators	Comments
	forum_topic_id	Ling		
	problem_id	Long		
	has_incidents	Boolean		
	due_at	DateTime		
	tags	List		
	via	Struct		
	custom_fields	List		
	satisfaction_rating	Struct		
	sharing_agreement_ids	List		
	followup_ids	List		
	via_followup_source_id	Long		
	ticket_form_id	Long		
	brand_id	Long		
	allow_channelback	Boolean		
	allow_attachments	Boolean		
	is_public	Boolean		

Entity	Field	Data type	Supported operators	Comments
	from_messaging_channel	Boolean		
	created_at	DateTime		
	updated_at	DateTime	EQUAL_TO	
	assignee_email	String		
	attribute_value_ids	List		
	collaborators	List		
	comment	Struct		
	custom_status_id	Long		
	email_ccs	Struct		
	followers	Struct		
	macro_id	Long		
	macros_ids	List		
	metadata	Struct		
	safe_update	Boolean		
	updated_stamp	DateTime		
	via_id	Long		
	voice_comment	Struct		

Entity	Field	Data type	Supported operators	Comments
	DML_STATUS	String		A user-defined field used to track the created, updated and deleted status of the record.
trigger-category	url	String		
	id	String		
	name	String		
	updated_at	DateTime		
	created_at	DateTime		
	position	Integer		
trigger	url	String		
	id	Long		
	title	String		
	active	Boolean	EQUAL_TO	
	updated_at	DateTime		
	created_at	DateTime		
	default	Boolean		
	actions	List		
	conditions	Struct		

Entity	Field	Data type	Supported operators	Comments
	description	String		
	position	Integer		
	raw_title	String		
	category_id	String	EQUAL_TO	
users	url	String		
	id	Long		
	external_id	String	EQUAL_TO	
	email	String		
	active	Boolean		
	alias	String		
	chat_only	Boolean		
	custom_roll_id	Long		
	roll_type	Integer		
	details	String		
	last_login_at	DateTime		
	locale	String		
	locale_id	Integer		
	moderator	Boolean		
	notes	String		
	name	String		

Entity	Field	Data type	Supported operators	Comments
	only_private_comments	Boolean		
	organization_id	Long		
	default_group_id	Long		
	phone	String		
	photo	Struct		
	remote_photo_url	String		
	restricted_agent	Boolean		
	role	String	EQUAL_TO	
	shared	Boolean		
	shared_agent	Boolean		
	tag	List		
	signature	String		
	suspended	Boolean		
	ticket_restriction	String		
	time_zone	String		
	iana_time_zone			
	two_factor_auth_enabled			

Entity	Field	Data type	Supported operators	Comments
	user_fields			
	verified	Boolean		
	report_csv	Boolean		
	created_at	DateTime		
	updated_at	DateTime	EQUAL_TO	
	permission_set	Long	EQUAL_TO	
	shared_phone_number	Boolean		
	DML_STATUS	String		A user-defined field used to track the created, updated and deleted status of the record.
view	url	String		
	id	Long		
	title	String		
	active	Boolean	EQUAL_TO	
	updated_at	DateTime		
	created_at	DateTime		
	default	Boolean		
	position	Integer		

Entity	Field	Data type	Supported operators	Comments
	description	String		
	execution	Struct		
	restriction	Struct		
	raw_title	String		
	conditions	Struct		
	access	String	EQUAL_TO	
	group_id	Long	EQUAL_TO	

Note

Struct and List data types are converted to String data type in the response of the connector.

Partitioning queries

Partitions are not supported in Zendesk.

Zendesk connection options

The following are connection options for Zendesk:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your Object in Zendesk.
- `API_VERSION`(String) - (Required) Used for Read. Zendesk Rest API version you want to use. For example: v2.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object. For example: id, name, url, created_at
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format. For example: group_id = 100

- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query. For example: "SELECT id,url FROM users WHERE role=\"end-user\""
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query. Default field is `update_at` for entities supporting the incremental export API (`created_at` for `ticket-events` and `time` for `ticket-metric-events`).
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field. Optional; this option will be handled by the connector if not provided in the job option. Default value - "2024-05-01T20:55:02.000Z
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read. Optional; this option will be handled by the connector if not provided in the job option. Default value : 1.
- `IMPORT_DELETED_RECORDS`(String) - Default: FALSE. Used for read. To get the delete records while querying.
- `ACCESS_TOKEN` - Access token to be used in the request.
- `INSTANCE_URL` - URL of the instance where the user wants to run the operations. For example : https://{subdomain}/.zendesk.com

Limitations

The following are limitations of the Zendesk connector:

- Offset-based pagination limits the number of pages that can be fetched to 100, but it not recommended as the total number of records that can be fetched is 10,000. However, the cursor-based pagination that is implemented for the Zendesk connector overcomes this limitation. Only the EQUAL_TO filter operator is supported through the Zendesk API.

  Because of this limitation, partitioning is not supported for the Zendesk connector.
- For the "Ticket Event" entity the Rate Limit is 10 requests per minute. While running a AWS Glue ETL job you may receive a 429 (too many requests) error.

Connecting to Zoho CRM

Zoho CRM acts as a single repository to bring sales, marketing, and customer support activities together, and streamline process, policy, and people in one platform. Zoho CRM can be easily customized to meet the specific needs of any business type and size.

Zoho CRM's developer platform offers the right mix of low-code and pro-code tools for businesses/ enterprises to automate work, integrate data across enterprise stack, and create custom solutions for web and mobile.

Topics

- AWS Glue support for Zoho CRM
- Policies containing the API operations for creating and using connections
- Configuring Zoho CRM
- Configuring Zoho CRM connections
- Reading from Zoho CRM entities
- Zoho CRM connection options
- Limitations and notes for Zoho CRM connector

AWS Glue support for Zoho CRM

AWS Glue supports Zoho CRM as follows:

Supported as a source?

Yes – Sync and Async. You can use AWS Glue ETL jobs to query data from Zoho CRM.

Supported as a target?

No.

Supported Zoho CRM API versions

The following Zoho CRM API versions are supported:

- v7

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Zoho CRM

Before you can use AWS Glue to transfer data from Zoho CRM, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- You have a Zoho CRM account.
- Your Zoho CRM account is enabled for API access.
- You have a registered API client under the API Console to obtain OAuth Credentials.

Configuring Zoho CRM connections

The grant type determines how AWS Glue communicates with Zoho CRM to request access to your data. Your choice affects the requirements that you must meet before you create the connection. Zoho CRM supports only the AUTHORIZATION_CODE grant type for OAuth 2.0.

- This grant type is considered "three-legged" OAuth as it relies on redirecting users to a third-party authorization server to authenticate the user. It is used when creating connections via the AWS Glue console. The AWS Glue console will redirect the user to Zoho CRM where the user must login and allow Glue the requested permissions to access their Zoho CRM instance.
- Users may still opt to create their own connected app in Zoho CRM and provide their own client ID, Auth URL, Token URL, and Instance URL when creating connections through the AWS Glue console. In this scenario, they will still be redirected to Zoho CRM to login and authorize AWS Glue to access their resources.
- This grant type results in a refresh token and access token. The access token will remain valid for one hour, and may be refreshed automatically without user interaction using the refresh token.
- For public Zoho CRM documentation on creating a connected app for Authorization Code OAuth flow, see Authentication.

To configure a Zoho CRM connection:

1. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
   a. When selecting a **Connection type**, select Zoho CRM.
   b. Provide the `INSTANCE_URL` of the Zoho CRM instance you want to connect to.
   c. Provide the user client application client ID.
   d. Select the appropriate **Auth URL** from the dropdown.
   e. Select the appropriate **Token URL** from the dropdown.

f. Select the AWS IAM role which AWS Glue can assume and has permissions for following actions:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

g. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.

h. Select the network options if you want to use your network.

2. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.
3. In your AWS Glue job configuration, provide `connectionName` as an **Additional network connection**.

Reading from Zoho CRM entities

Prerequisite

Zoho CRM objects you would like to read from. You will need the object name.

Supported entities for Sync source:

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Product	Yes	Yes	Yes	Yes	Yes
Quote	Yes	Yes	Yes	Yes	Yes
Purchase Order	Yes	Yes	Yes	Yes	Yes
Solution	Yes	Yes	Yes	Yes	Yes
Call	Yes	Yes	Yes	Yes	Yes
Task	Yes	Yes	Yes	Yes	Yes
Event	Yes	Yes	Yes	Yes	Yes
Invoice	Yes	Yes	Yes	Yes	Yes
Account	Yes	Yes	Yes	Yes	Yes
Contact	Yes	Yes	Yes	Yes	Yes
Vendor	Yes	Yes	Yes	Yes	Yes
Campaign	Yes	Yes	Yes	Yes	Yes
Deal	Yes	Yes	Yes	Yes	Yes
Lead	Yes	Yes	Yes	Yes	Yes
Custom Module	Yes	Yes	Yes	Yes	Yes
Sales Order	Yes	Yes	Yes	Yes	Yes
Price Books	Yes	Yes	Yes	Yes	Yes
Case	Yes	Yes	Yes	Yes	Yes

Example:

```
zoho_read = glueContext.create_dynamic_frame.from_options(
    connection_type="ZOHO",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "entityName",
        "API_VERSION": "v7",
        "INSTANCE_URL": "https://www.zohoapis.in/"
    }
```

Supported entities for Async source:

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Product	Yes	No	No	Yes	No
Quote	Yes	No	No	Yes	No
Purchase Order	Yes	No	No	Yes	No
Solution	Yes	No	No	Yes	No
Call	Yes	No	No	Yes	No
Task	Yes	No	No	Yes	No
Event	Yes	No	No	Yes	No
Invoice	Yes	No	No	Yes	No
Account	Yes	No	No	Yes	No
Contact	Yes	No	No	Yes	No
Vendor	Yes	No	No	Yes	No
Campaign	Yes	No	No	Yes	No

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Deal	Yes	No	No	Yes	No
Lead	Yes	No	No	Yes	No
Custom Module	Yes	No	No	Yes	No
Sales Order	Yes	No	No	Yes	No
Price Books	Yes	No	No	Yes	No
Case	Yes	No	No	Yes	No

Example:

```
zoho_read = glueContext.create_dynamic_frame.from_options(
    connection_type="ZOHO",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "entityName",
        "API_VERSION": "v7",
        "INSTANCE_URL": "https://www.zohoapis.in/",
        "TRANSFER_MODE": "ASYNC"
    }
```

Zoho CRM field details:

Zoho CRM provides endpoints to fetch metadata dynamically for supported entities. Therefore, operator support is captured at the datatype level.

Entity	Data type	Supported operators
Zoho entities (all entities)	Integer	!=, =, <, <=, >, >=, BETWEEN
	String	Like, =, !=
	BigInteger	!=, =, <, <=, >, >=, BETWEEN

Entity	Data type	Supported operators
	Boolean	=
	Double	!=, =, <, <=, >, >=, BETWEEN
	BigDecimal	!=, =, <, <=, >, >=, BETWEEN
	Date	!=, =, <, <=, >, >=, BETWEEN
	DateTime	!=, =, <, <=, >, >=, BETWEEN
	Struct	N/A
	List	N/A

Partitioning queries

Partitioning is not supported in Async mode.

Filter-based partitioning (Sync mode):

You can provide the additional Spark options `PARTITION_FIELD`, `LOWER_BOUND`, `UPPER_BOUND`, and `NUM_PARTITIONS` if you want to utilize concurrency in Spark. With these parameters, the original query would be split into `NUM_PARTITIONS` number of sub-queries that can be executed by Spark tasks concurrently.

- `PARTITION_FIELD`: the name of the field to be used to partition the query.
- `LOWER_BOUND`: an **inclusive** lower bound value of the chosen partition field.

  For the Datetime field, we accept the Spark timestamp format used in Spark SQL queries.

  Examples of valid value:

  ```
  "2024-09-30T01:01:01.000Z"
  ```
- `UPPER_BOUND`: an **exclusive** upper bound value of the chosen partition field.
- `NUM_PARTITIONS`: the number of partitions.

Example:

```
zoho_read = glueContext.create_dynamic_frame.from_options(
    connection_type="zohocrm",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "entityName",
        "API_VERSION": "v7",
        "PARTITION_FIELD": "Created_Time"
        "LOWER_BOUND": "2022-01-01T01:01:01.000Z"
        "UPPER_BOUND": "2024-01-01T01:01:01.000Z"
        "NUM_PARTITIONS": "10"
    }
```

Zoho CRM connection options

The following are connection options for Zoho CRM:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of your object in Zoho CRM.
- `API_VERSION`(String) - (Required) Used for Read. Zoho CRM Rest API version you want to use.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. Columns you want to select for the object.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.
- `PARTITION_FIELD`(String) - Used for Read. Field to be used to partition query.
- `LOWER_BOUND`(String)- Used for Read. An inclusive lower bound value of the chosen partition field.
- `UPPER_BOUND`(String) - Used for Read. An exclusive upper bound value of the chosen partition field.
- `NUM_PARTITIONS`(Integer) - Default: 1. Used for Read. Number of partitions for read.
- `INSTANCE_URL`(String) - (Required) Used for Read. A valid Zoho CRM instance URL.
- `TRANSFER_MODE`(String) - Used to indicate whether the query should be run on Async mode.

Limitations and notes for Zoho CRM connector

The following are limitations or notes for the Zoho CRM connector:

- With API version v7, you can fetch a maximum of 100,000 records. See the Zoho documentation .
- For the Event entity, the label "Meeting" is displayed as mentioned in the Zoho documentation.
- For Select All functionality:
    - You can fetch a maximum of 50 fields from SaaS for both the GET and POST call.
    - If you want to have data for some specific field that does not belong in the first 50 fields, you will need to manually provide the list of selected fields.
    - If more than 50 fields are selected, any fields beyond the 50 fields will be trimmed and will contain null data in Amazon S3.
    - In case of a filter expression, if the user-provided list of 50 fields does not include "id" and "Created_Time," a custom exception will be raised to prompt the user to include these fields.
- Filter operators may vary from field-to-field despite of having the same data type. Therefore, you must manually specify a different operator for any field that triggers an error in the SaaS platform.
- For Sort By functionality:
    - Data can only be sorted by a single field without a filter expression, whereas data can be sorted by multiple fields when a filter expression is applied.
    - If no sort order is specified for the selected field, the data will be retrieved in ascending order by default.
- The supported regions for the Zoho CRM connector are US, Europe, India, Australia and Japan.
- Async read functionality Limitations:
    - Limit order by and partitioning is not supported in the Async mode.
    - In the Async mode we can transfer data up to 500 pages with 200,000 records per page.
    - For a one-minute interval, only 10 requests are allowed for download. When you exceed the download limit, the system returns an HTTP 429 error and pauses all download requests for one minute before processing can resume.
    - After completing the bulk job, you can access the downloadable file only for a period of one day. After that, you cannot access the file via endpoints.
    - A maximum of 200 select fields can be given via an endpoint. If you specify more than 200 select fields in an endpoint, the system will automatically export all available fields for that module.
    - External fields created in any module are not supported in Bulk Read APIs.
    - Sorting and `Group_by` clauses are not supported via this API endpoint.

- The values of the fields with sensitive health data will be retrieved only when the **Restrict Data access through API** option in the compliance settings is **disabled**. If the option is enabled, the value will be **empty** in the result.
- Filtration/Criteria Limits
  - The maximum number of criteria that can be used in a query is 25.
  - Filtration/Criteria on multiline text fields is not supported.

Connecting to Zoom Meetings

Zoom Meetings is a cloud-based video conferencing platform that can be used for video conferencing meetings, audio conferencing, webinars, meeting recordings, and live chat.

Topics

- AWS Glue support for Zoom Meetings
- Policies containing the API operations for creating and using connections
- Configuring Zoom Meetings
- Configuring the Zoom Meetings client app
- Configuring Zoom Meetings connections
- Reading from Zoom Meetings entities
- Zoom Meetings connection options
- Zoom Meetings limitations

AWS Glue support for Zoom Meetings

AWS Glue supports Zoom Meetings as follows:

Supported as a source?

Yes. You can use AWS Glue ETL jobs to query data from Zoom Meetings.

Supported as a target?

No.

Supported Zoom Meetings API versions

The following Zoom Meetings API versions are supported:

- v2

Policies containing the API operations for creating and using connections

The following sample policy describes the required AWS IAM permissions for creating and using connections. If you are creating a new role, create a policy that contains the following:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListConnectionTypes",
        "glue:DescribeConnectionType",
        "glue:RefreshOAuth2Tokens",
        "glue:ListEntities",
        "glue:DescribeEntity"
      ],
      "Resource": "*"
    }
  ]
}
```

If you don't want to use the above method, alternatively use the following managed IAM policies:

- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.
- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.

Configuring Zoom Meetings

Before you can use AWS Glue to transfer data from Zoom Meetings, you must meet these requirements:

Minimum requirements

The following are minimum requirements:

- You have a Zoom Meetings account.
- Your Zoom account is enabled for API access.
- You have created an OAuth2 app in your Zoom Meetings account. This integration provides the credentials that AWS Glue uses to access your data securely when it makes authenticated calls to your account. For more information, see the section called “Configuring the Zoom Meetings client app”.

If you meet these requirements, you’re ready to connect AWS Glue to your Zoom Meetings account. For typical connections, you don't need do anything else in Zoom Meetings.

Configuring the Zoom Meetings client app

1. Log into the Zoom App Marketplace.
2. Choose **Develop** > **Build App**.
3. Choose **General App** for an OAuth 2.0 based app.
4. On the **Basic Info** page, add or update information about the app such as the app's name, how the app is managed, app credentials, and OAuth information.
5. In the **Select how the app is managed** section, confirm how you want your app to be managed:
    a. **Admin-managed**: Account admins add and manage the app
    b. **User-managed**: Individual users add and manage the app. The app has access to only the user's authorized data.
6. **App Credentials**: The build flow automatically generates app credentials (client ID and client secret) for your app.
7. In the OAuth Information section, set up OAuth for your app.
    a. **OAuth redirect URL** (required): Enter your redirect URL or endpoint to set up OAuth between your app and Zoom.
    b. **Use Strict Mode URL** (optional)

   c. **Subdomain check** (optional)

   d. **OAuth allow lists** (required): Add any unique URLs that Zoom should allow as valid redirects for your OAuth flows.

8. On the **Scopes** page, select the Zoom API methods your app is allowed to call. The scopes define which information and capabilities are available to your user. Select the following granular scopes:

   - user:read:list_users:admin
   - zoom_rooms:read:list_rooms:admin
   - group:read:list_members:admin
   - group:read:administrator:admin
   - group:read:list_groups:admin
   - report:read:admin
   - role:read:list_roles, role:read:list_roles:admin

   Once the scopes are added choose **Continue** and the app is ready to use.

For more information about OAuth 2.0 setup see Integrations (OAuth apps).

Configuring Zoom Meetings connections

Zoom Meetings supports the AUTHORIZATION_CODE grant type for OAuth2. The grant type determines how AWS Glue communicates with Zoom Meetings to request access to your data.

- This grant type is considered "three-legged" OAuth as it relies on redirecting users to a third-party authorization server to authenticate the user. It is used when creating connections via the AWS Glue console. The user creating a connection need to provide OAuth related information like Client ID and Client Secret for their Zoom Meetings client application. The AWS Glue console will redirect the user to Zoom where the user must login and allow AWS Glue the requested permissions to access their Zoom Meetings instance.
- Users may still opt to create their own connected app in Zoom Meetings and provide their own client ID and client secret when creating connections through the AWS Glue console. In this scenario, they will still be redirected to Zoom Meetings to login and authorize AWS Glue to access their resources.
- This grant type results in a refresh token and access token. The access token is short lived, and may be refreshed automatically without user interaction using the refresh token.

- For public Zoom Meetings documentation on creating a connected app for Authorization Code OAuth flow, see Using OAuth 2.0.

To configure a Zoom Meetings connection:

1. In AWS Secrets Manager, create a secret with the following details:
    a. For the customer managed connected app, the Secret should contain the connected app Consumer Secret with `USER_MANAGED_CLIENT_APPLICATION_CLIENT_SECRET` as key.
    b. Note: you must create a secret for your connections in AWS Glue.

1. In AWS Glue Glue Studio, create a connection under **Data Connections** by following the steps below:
    a. When selecting a **Connection type**, select Zoom Meetings.
    b. Provide the Zoom Meetings environment you want to connect to.
    c. Select the AWS IAM role which AWS Glue can assume and has permissions for following actions:

       JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:DescribeSecret",
        "secretsmanager:GetSecretValue",
        "secretsmanager:PutSecretValue",
        "ec2:CreateNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DeleteNetworkInterface"
      ],
      "Resource": "*"
    }
  ]
}
```

    d. Select the `secretName` which you want to use for this connection in AWS Glue to put the tokens.

   e. Select the network options if you want to use your network.

2. Grant the IAM role associated with your AWS Glue job permission to read `secretName`.

Reading from Zoom Meetings entities

Prerequisite

A Zoom Meetings object you would like to read from. You will need the object namem such as `Group` or `Zoom Rooms`.

Supported entities for source:

Entity	Can be filtered	Supports limit	Supports Order by	Supports Select *	Supports partitioning
Zoom Rooms	No	Yes	No	Yes	No
Group	No	No	No	Yes	No
Group Member	Yes	Yes	No	Yes	No
Group Admin	No	Yes	No	Yes	No
Report (daily)	Yes	No	No	Yes	No
Roles	No	No	No	Yes	No
Users	Yes	Yes	No	Yes	No

Example:

```
zoom_read = glueContext.create_dynamic_frame.from_options(
    connection_type="zoom",
    connection_options={
        "connectionName": "connectionName",
        "ENTITY_NAME": "organization",
        "API_VERSION": "v2"
    }
)
```

Zoom Meetings entity and field details:

Zoom Meetings dynamically loads the available fields under the selected entity. Depending on the data type of the field, it supports the following filter operators.

Entity	Field	Data type	Supported operators	
Zoom Room	status	String	=	
	type	String	=	
	unassigned_rooms	Boolean	=	
	location_id	String	=	
	room_id	String	N/A	
	activation_code	String	N/A	
	id	String	N/A	
	name	String	N/A	
	tag_ids	String	N/A	
	query_name	String	N/A	
Daily Report	month	Date	=	
	date	Date	N/A	
	meeting_minutes	Integer	N/A	
	meetings	Integer	N/A	
	new_users	Integer	N/A	

Entity	Field	Data type	Supported operators	
	participants	Integer	N/A	
	group_id	String	N/A	
User	created_at	DateTime	N/A	
	dept	String	N/A	
	email	String	N/A	
	employee_unique_id	String	N/A	
	first_name	String	N/A	
	group_ids	List	N/A	
	host_key	String	N/A	
	id	String	N/A	
	im_group_ids	String	N/A	
	last_client_version	String	N/A	
	last_login_time	DateTime	N/A	
	last_name	String	N/A	
	plan_united_type	String	N/A	
	custom_attributes	List	N/A	
	pmi	BigInteger	N/A	

Entity	Field	Data type	Supported operators	
	role_id	String	=	
	status	String	=	
	timezone	String	N/A	
	type	Integer	N/A	
	verified	Integer	N/A	
	user_created_at	DateTime	N/A	
	display_name	String	N/A	
	phone_number	String	N/A	
	language	String	N/A	
	license	String	=	
Group	id	String	N/A	
	name	String	N/A	
	total_members	Integer	N/A	
Group Member	email	String	N/A	
	first_name	String	N/A	
	id	String	N/A	
	last_name	String	N/A	
	type	Integer	N/A	
	primary_group	Boolean	N/A	
	member_id	String	N/A	

Entity	Field	Data type	Supported operators	
Group Admin	id	String	N/A	
	email	String	N/A	
	name	String	N/A	
role	description	String	N/A	
	id	String	N/A	
	name	String	N/A	
	total_members	Integer	N/A	
	type	String	=	

Partitioning queries

Zoom Meetings doesn't support filter-based partitioning or record-based partitioning.

Zoom Meetings connection options

The following are connection options for Zoom Meetings:

- `ENTITY_NAME`(String) - (Required) Used for Read. The name of the Zoom Meetings entity. For example, `group`.
- `API_VERSION`(String) - (Required) Used for Read. Zoom Meetings Rest API version you want to use. The value will be `v2`, as Zoom Meetings currently supports only version v2.
- `SELECTED_FIELDS`(List<String>) - Default: empty(SELECT *). Used for Read. A comma-separated list of columns you want to select for the selected entity.
- `FILTER_PREDICATE`(String) - Default: empty. Used for Read. It should be in the Spark SQL format.
- `QUERY`(String) - Default: empty. Used for Read. Full Spark SQL query.

Zoom Meetings limitations

The following are limitations or notes for Zoom Meetings:

- Zoom Meetings does not support orderby.
- Zoom Meetings does not support filter-based partitioning because there is no field that can satisfy the required criteria.
- Zoom Meetings does not support record-based partitioning because the pagination limit and offset-based pagination is not supported.

Adding a JDBC connection using your own JDBC drivers

You can use your own JDBC driver when using a JDBC connection. When the default driver utilized by the AWS Glue crawler is unable to connect to a database, you can use your own JDBC Driver. For example, if you want to use SHA-256 with your Postgres database, and older postgres drivers do not support this, you can use your own JDBC driver.

Supported datasources

Supported datasources	Unsupported datasources
MySQL	Snowflake
Postgres	
Oracle	
Redshift	
SQL Server	
Aurora*	

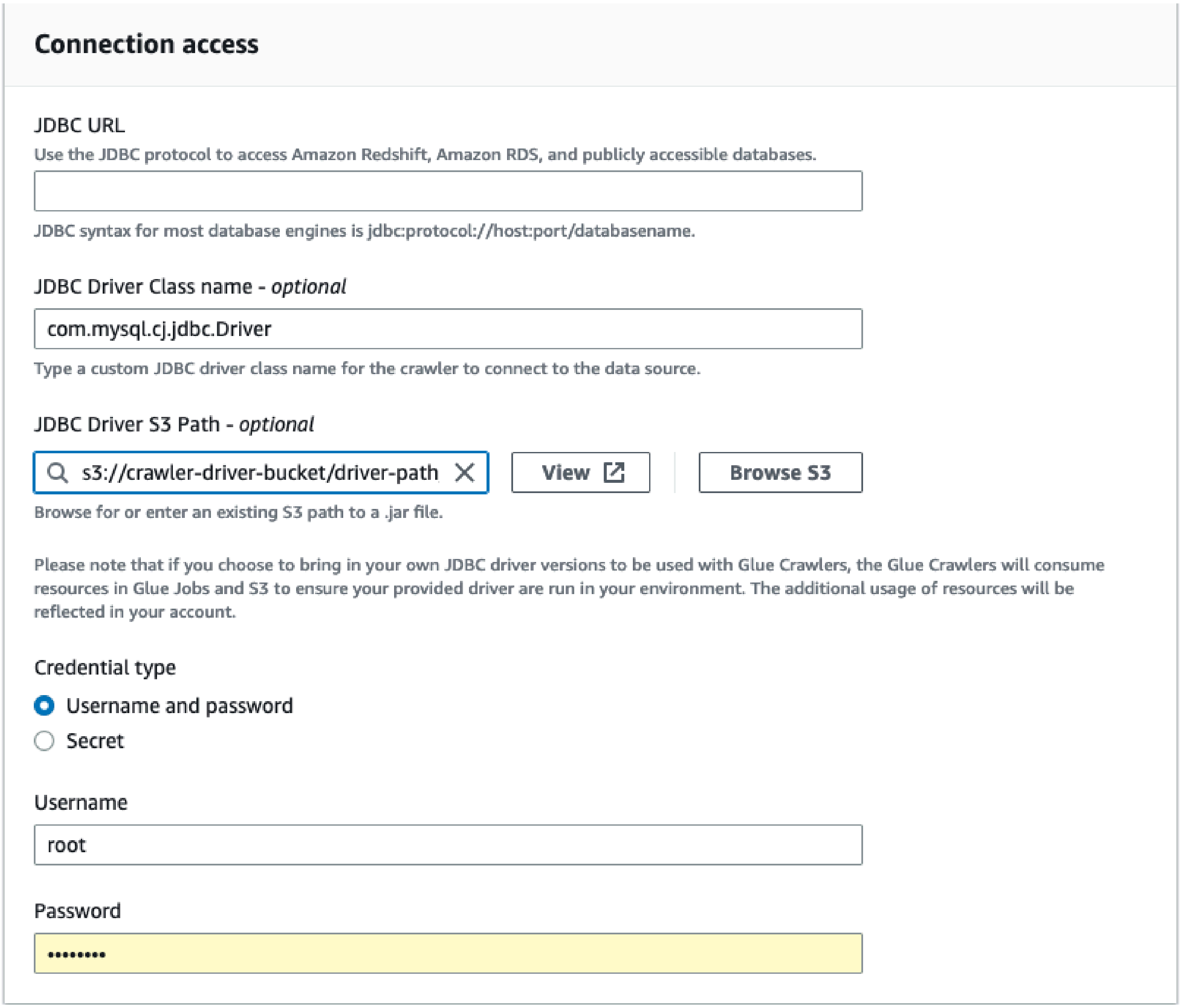
*Supported if the native JDBC driver is being used. Not all driver features can be leveraged.

Adding a JDBC driver to a JDBC connection

> **Note**
>
> If you choose to bring in your own JDBC driver versions, AWS Glue crawlers will consume resources in AWS Glue jobs and Amazon S3 buckets to ensure your provided driver are run in your environment. The additional usage of resources will be reflected in your account. The cost for AWS Glue crawlers and jobs is under the AWS Glue category in billing. Additionally, providing your own JDBC driver does not mean that the crawler is able to leverage all of the driver's features.

To add your own JDBC driver to a JDBC connection:

1. Add the JDBC driver file to an Amazon S3 location. You can create a bucket and/or folder or use an existing bucket and/or folder.
2. In the AWS Glue console, choose **Connections** in the left-hand menu under **Data Catalog**, then create a new connection.
3. Complete the fields for **Connection properties** and choose JDBC for **Connection type**.
4. In **Connection access**, enter the **JDBC URL** and **JDBC Driver Class name –** *optional*. The driver class name must be for a datasource supported by AWS Glue crawlers.



Connection access

JDBC URL

Use the JDBC protocol to access Amazon Redshift, Amazon RDS, and publicly accessible databases.

JDBC syntax for most database engines is jdbc:protocol://host:port/databasename.

JDBC Driver Class name - *optional*

com.mysql.cj.jdbc.Driver

Type a custom JDBC driver class name for the crawler to connect to the data source.

JDBC Driver S3 Path - *optional*

s3://crawler-driver-bucket/driver-path

Browse for or enter an existing S3 path to a .jar file.

Please note that if you choose to bring in your own JDBC driver versions to be used with Glue Crawlers, the Glue Crawlers will consume resources in Glue Jobs and S3 to ensure your provided driver are run in your environment. The additional usage of resources will be reflected in your account.

Credential type

● Username and password

○ Secret

Username

root

Password

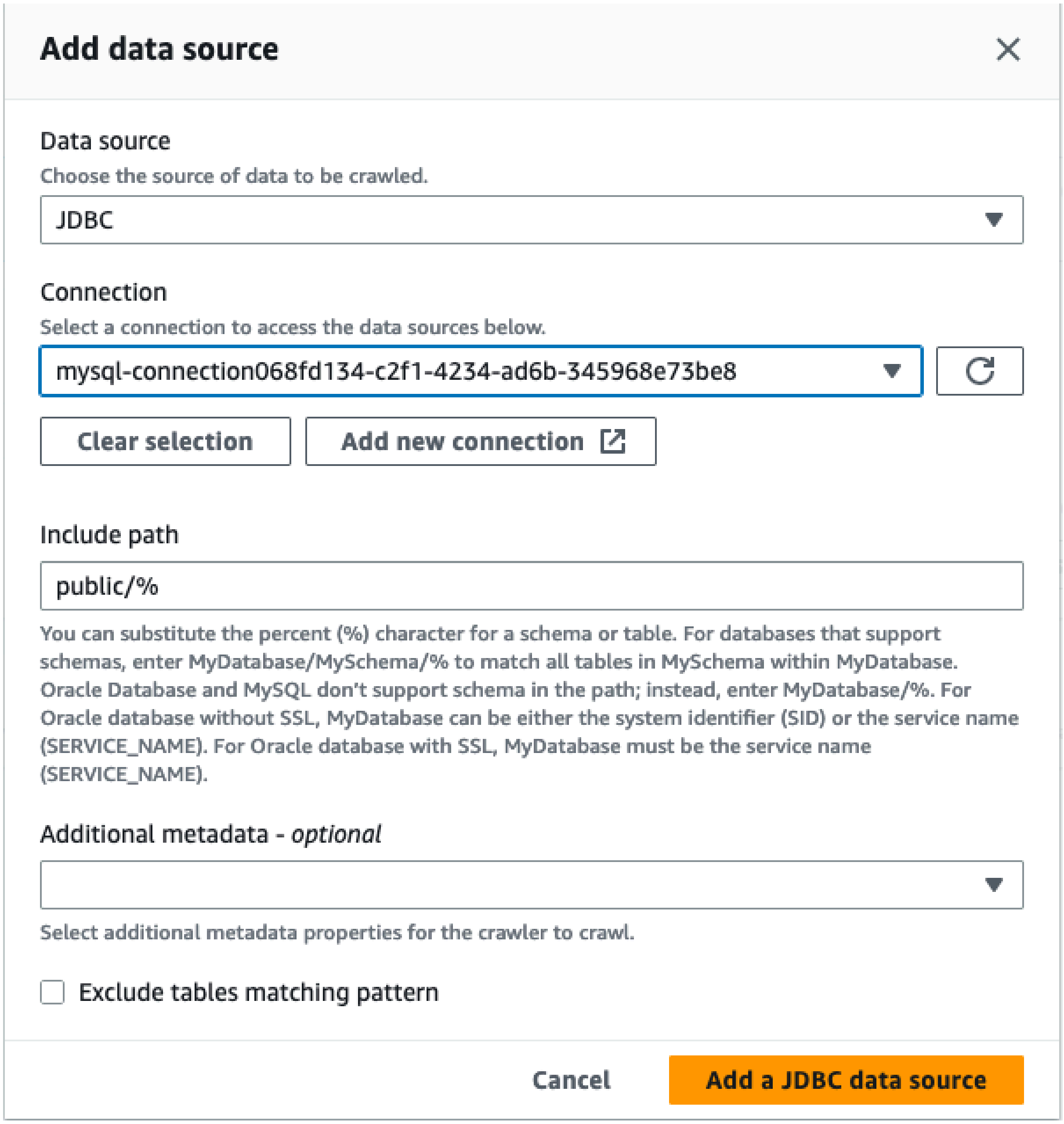
••••••••

5. Choose the Amazon S3 path where the JDBC driver is located in the **JDBC Driver Amazon S3 Path** – *optional* field.
6. Complete the fields for Credential type if entering a username and password or secret. When complete, choose **Create connection**.

> **Note**
>
> Testing connection is not supported currently. When crawling the data source with a JDBC driver you provided, the crawler skips this step.

7. Add the newly created connection to a crawler. In the AWS Glue console, choose **Crawlers** in the left-hand menu under **Data Catalog**, then create a new crawler.

8. In the **Add crawler** wizard, in Step 2 choose **Add a data source**.



Data source

Choose the source of data to be crawled.

JDBC

Connection

Select a connection to access the data sources below.

mysql-connection068fd134-c2f1-4234-ad6b-345968e73be8

Include path

public/%

You can substitute the percent (%) character for a schema or table. For databases that support schemas, enter MyDatabase/MySchema/% to match all tables in MySchema within MyDatabase. Oracle Database and MySQL don't support schema in the path; instead, enter MyDatabase/%. For Oracle database without SSL, MyDatabase can be either the system identifier (SID) or the service name (SERVICE_NAME). For Oracle database with SSL, MyDatabase must be the service name (SERVICE_NAME).

Additional metadata - *optional*

Select additional metadata properties for the crawler to crawl.

☐ Exclude tables matching pattern

9. Choose **JDBC** as the data source and choose the the connection that was created in the previous steps. Complete

10. In order to use your own JDBC driver with a AWS Glue crawler, add the following permissions to the role used by the crawler:

    - Grant permissions for the following job actions: `CreateJob`, `DeleteJob`, `GetJob`, `GetJobRun`, `StartJobRun`.
    - Grant permissions for IAM actions: `iam:PassRole`

- Grant permissions for Amazon S3 actions: `s3:DeleteObjects`, `s3:GetObject`, `s3:ListBucket`, `s3:PutObject`.
- Grant service principal access to bucket/folder in the IAM policy.

Example IAM policy:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "VisualEditor0",
      "Effect": "Allow",
      "Action": [
        "s3:PutObject",
        "s3:GetObject",
        "s3:ListBucket",
        "s3:DeleteObject"
      ],
      "Resource": [
        "arn:aws:s3:::amzn-s3-demo-bucket/driver-parent-folder/driver.jar",
        "arn:aws:s3:::amzn-s3-demo-bucket"
      ]
    }
  ]
}
```

The AWS Glue crawler creates two folders: _glue_job_crawler and _crawler.

If the driver jar is located in the `s3://amzn-s3-demo-bucket/driver.jar"` folder, add the following resources:

```
"Resource": [
                "arn:aws:s3:::amzn-s3-demo-bucket/_glue_job_crawler/*",
     "arn:aws:s3:::amzn-s3-demo-bucket/_crawler/*"
           ]
```

If the driver jar is located in the `s3://amzn-s3-demo-bucket/tmp/driver/subfolder/driver.jar"`folder, add the following resources:

```
"Resource": [
                "arn:aws:s3:::amzn-s3-demo-bucket/tmp/_glue_job_crawler/*",
    "arn:aws:s3:::amzn-s3-demo-bucket/tmp/_crawler/*"
            ]
```

11. If you are using a VPC, you must allow access to the AWS Glue endpoint by creating the interface endpoint and add it to your route table. For more information, see Creating an interface VPC endpoint for AWS Glue
12. If you are using encryption in your Data Catalog, create the AWS KMS interface endpoint and add it to your route table. For more information, see Creating a VPC endpoint for AWS KMS.

Using custom connectors and connections with AWS Glue Studio

AWS Glue provides built-in support for the most commonly used data stores (such as Amazon Redshift, Amazon Aurora, Microsoft SQL Server, MySQL, MongoDB, and PostgreSQL) using JDBC connections. AWS Glue also allows you to use custom JDBC drivers in your extract, transform, and load (ETL) jobs. For data stores that are not natively supported, such as SaaS applications, you can use connectors.

A *connector* is an optional code package that assists with accessing data stores in AWS Glue Studio. You can subscribe to several connectors offered in AWS Marketplace.

When creating ETL jobs, you can use a natively supported data store, a connector from AWS Marketplace, or your own custom connectors. If you use a connector, you must first create a connection for the connector. A *connection* contains the properties that are required to connect to a particular data store. You use the connection with your data sources and data targets in the ETL job. Connectors and connections work together to facilitate access to the data stores.

The following connections are available when creating connections for connectors:

- **Amazon Aurora** – a scalable, high-performance relational database engine with built-in security, backup and restore, and in-memory acceleration.

- **Amazon DocumentDB** – a scalable, highly available, and fully managed document database service that supports MongoDB and SQL APIs.
- **Amazon Redshift** – a scalable, highly available, and fully managed document database service that supports MongoDB and SQL APIs.
- **Azure SQL** – a cloud-based relational database service from Microsoft Azure that provides scalable, reliable, and secure data storage and management capabilities.
- **Cosmos DB** – a globally distributed cloud database service from Microsoft Azure that provides scalable, high-performance data storage and querying capabilities.
- **Google BigQuery** – a serverless cloud data warehouse for running fast SQL queries on large datasets.
- **JDBC** – a relational database management system (RDBMS) that uses a Java API for connecting and interacting with data connections.
- **Kafka** – an open-source stream processing platform used for real-time data streaming and messaging.
- **MariaDB** – a community-developed fork of MySQL that offers enhanced performance, scalability, and features.
- **MongoDB** – a cross-platform document-oriented database that provides high scalability, flexibility, and performance.
- **MongoDB Atlas** – a cloud-based database as a service (DBaaS) offering from MongoDB that simplifies the management and scaling of MongoDB deployments.
- **Microsoft SQL Server** – a relational database management system (RDBMS) from Microsoft that provides robust data storage, analysis, and reporting capabilities.
- **Mixpanel** – an analytics platform that helps businesses analyze how users interact with their websites, mobile applications, and other digital products.
- **MySQL** – an open-source relational database management system (RDBMS) that is widely used in web applications and is known for its reliability and scalability.
- **Network** – a network data source represents a network-accessible resource or service that can be accessed by a data integration platform.
- **OpenSearch** – an OpenSearch data source is an application that OpenSearch can connect to and ingest data from.
- **Oracle** – a relational database management system (RDBMS) from Oracle Corporation that provides robust data storage, analysis, and reporting capabilities.

- **PostgreSQL** – an open-source relational database management system (RDBMS) that provides robust data storage, analysis, and reporting capabilities.
- **Salesforce** – Salesforce provides customer relationship management (CRM) software that help you with sales, customer service, e-commerce, and more. If you're a Salesforce user, you can connect AWS Glue to your Salesforce account. Then, you can use Salesforce as a data source or destination in your ETL jobs. Run these jobs to transfer data between Salesforce and AWS services or other supported applications.
- **SAP HANA** – an in-memory database and analytics platform that provides fast data processing, advanced analytics, and real-time data integration.
- **Snowflake** – a cloud-based data warehouse that provides scalable, high-performance data storage and analytics services.
- **Teradata** – a relational database management system (RDBMS) that provides high-performance data storage, analysis, and reporting capabilities.
- **Vertica** – a columnar-oriented analytical data warehouse designed for big data analytics that offers fast query performance, advanced analytics, and scalability.

Creating custom connectors

You can also build your own connector and then upload the connector code to AWS Glue Studio.

Custom connectors are integrated into AWS Glue Studio through the AWS Glue Spark runtime API. The AWS Glue Spark runtime allows you to plug in any connector that is compliant with the Spark, Athena, or JDBC interface. It allows you to pass in any connection option that is available with the custom connector.

You can encapsulate all your connection properties with AWS Glue Connections and supply the connection name to your ETL job. Integration with Data Catalog connections allows you to use the same connection properties across multiple calls in a single Spark application or across different applications.

You can specify additional options for the connection. The job script that AWS Glue Studio generates contains a `Datasource` entry that uses the connection to plug in your connector with the specified connection options. For example:

```
Datasource = glueContext.create_dynamic_frame.from_options(connection_type =
"custom.jdbc", connection_options = {"dbTable":"Account","connectionName":"my-custom-
jdbc-
connection"}, transformation_ctx = "DataSource0")
```

To add a custom connector to AWS Glue Studio

1. Create the code for your custom connector. For more information, see Developing custom connectors.
2. Add support for AWS Glue features to your connector. Here are some examples of these features and how they are used within the job script generated by AWS Glue Studio:

   - **Data type mapping** – Your connector can typecast the columns while reading them from the underlying data store. For example, a `dataTypeMapping` of `{"INTEGER":"STRING"}` converts all columns of type `Integer` to columns of type `String` when parsing the records and constructing the `DynamicFrame`. This helps users to cast columns to types of their choice.

     ```
     DataSource0 = glueContext.create_dynamic_frame.from_options(connection_type
     = "custom.jdbc", connection_options = {"dataTypeMapping":{"INTEGER":"STRING"}",
     connectionName":"test-connection-jdbc"}, transformation_ctx = "DataSource0")
     ```

   - **Partitioning for parallel reads** – AWS Glue allows parallel data reads from the data store by partitioning the data on a column. You must specify the partition column, the lower partition bound, the upper partition bound, and the number of partitions. This feature enables you to make use of data parallelism and multiple Spark executors allocated for the Spark application.

     ```
     DataSource0 = glueContext.create_dynamic_frame.from_options(connection_type
     = "custom.jdbc", connection_options = {"upperBound":"200","numPartitions":"4",
     "partitionColumn":"id","lowerBound":"0","connectionName":"test-connection-jdbc"},
     transformation_ctx = "DataSource0")
     ```

   - **Use AWS Secrets Manager for storing credentials** –The Data Catalog connection can also contain a `secretId` for a secret stored in AWS Secrets Manager. The AWS secret can securely store authentication and credentials information and provide it to AWS Glue at runtime. Alternatively, you can specify the `secretId` from the Spark script as follows:

     ```
     DataSource = glueContext.create_dynamic_frame.from_options(connection_type
     = "custom.jdbc", connection_options = {"connectionName":"test-connection-jdbc",
      "secretId"-> "my-secret-id"}, transformation_ctx = "DataSource0")
     ```

- **Filtering the source data with row predicates and column projections** – The AWS Glue Spark runtime also allows users to push down SQL queries to filter data at the source with row predicates and column projections. This allows your ETL job to load filtered data faster from data stores that support push-downs. An example SQL query pushed down to a JDBC data source is: `SELECT id, name, department FROM department WHERE id < 200.`

  ```
  DataSource = glueContext.create_dynamic_frame.from_options(connection_type =
  "custom.jdbc", connection_options = {"query":"SELECT id, name, department FROM
   department
  WHERE id < 200","connectionName":"test-connection-jdbc"}, transformation_ctx =
  "DataSource0")
  ```

- **Job bookmarks** – AWS Glue supports incremental loading of data from JDBC sources. AWS Glue keeps track of the last processed record from the data store, and processes new data records in the subsequent ETL job runs. Job bookmarks use the primary key as the default column for the bookmark key, provided that this column increases or decreases sequentially. For more information about job bookmarks, see Job Bookmarks in the *AWS Glue Developer Guide.*

  ```
  DataSource0 = glueContext.create_dynamic_frame.from_options(connection_type =
  "custom.jdbc", connection_options = {"jobBookmarkKeys":["empno"],
   "jobBookmarkKeysSortOrder"
  :"asc", "connectionName":"test-connection-jdbc"}, transformation_ctx =
   "DataSource0")
  ```

3. Package the custom connector as a JAR file and upload the file to Amazon S3.
4. Test your custom connector. For more information, see the instructions on GitHub at Glue Custom Connectors: Local Validation Tests Guide.
5. In the AWS Glue Studio console, choose **Connectors** in the console navigation pane.
6. On the **Connectors** page, choose **Create custom connector**.
7. On the **Create custom connector** page, enter the following information:
   - The path to the location of the custom code JAR file in Amazon S3.
   - A name for the connector that will be used by AWS Glue Studio.
   - Your connector type, which can be one of **JDBC**, **Spark**, or **Athena**.

  - The name of the entry point within your custom code that AWS Glue Studio calls to use the connector.
    - For JDBC connectors, this field should be the class name of your JDBC driver.
    - For Spark connectors, this field should be the fully qualified data source class name, or its alias, that you use when loading the Spark data source with the `format` operator.
  - (JDBC only) The base URL used by the JDBC connection for the data store.
  - (Optional) A description of the custom connector.
8. Choose **Create connector**.
9. From the **Connectors** page, create a connection that uses this connector, as described in [Creating connections for connectors](#).

Adding connectors to AWS Glue Studio

A connector is a piece of code that facilitates communication between your data store and AWS Glue. You can either subscribe to a connector offered in AWS Marketplace, or you can create your own custom connector.

Subscribing to AWS Marketplace connectors

AWS Glue Studio makes it easy to add connectors from AWS Marketplace.

To add a connector from AWS Marketplace to AWS Glue Studio

1. In the AWS Glue Studio console, choose **Connectors** in the console navigation pane.
2. On the **Connectors** page, choose **Go to AWS Marketplace**.
3. In AWS Marketplace, in **Featured products**, choose the connector you want to use. You can choose one of the featured connectors, or use search. You can search on the name or type of connector, and you can use options to refine the search results.

   If you want to use one of the featured connectors, choose **View product**. If you used search to locate a connector, then choose the name of the connector.
4. On the product page for the connector, use the tabs to view information about the connector. If you decide to purchase this connector, choose **Continue to Subscribe**.
5. Provide the payment information, and then choose **Continue to Configure**.
6. On the **Configure this software** page, choose the method of deployment and the version of the connector to use. Then choose **Continue to Launch**.

7. On the **Launch this software** page, you can review the **Usage Instructions** provided by the connector provider. When you're ready to continue, choose **Activate connection in AWS Glue Studio**.

   After a small amount of time, the console displays the **Create marketplace connection** page in AWS Glue Studio.

8. Create a connection that uses this connector, as described in Creating connections for connectors.

   Alternatively, you can choose **Activate connector only** to skip creating a connection at this time. You must create a connection at a later date before you can use the connector.

Creating connections for connectors

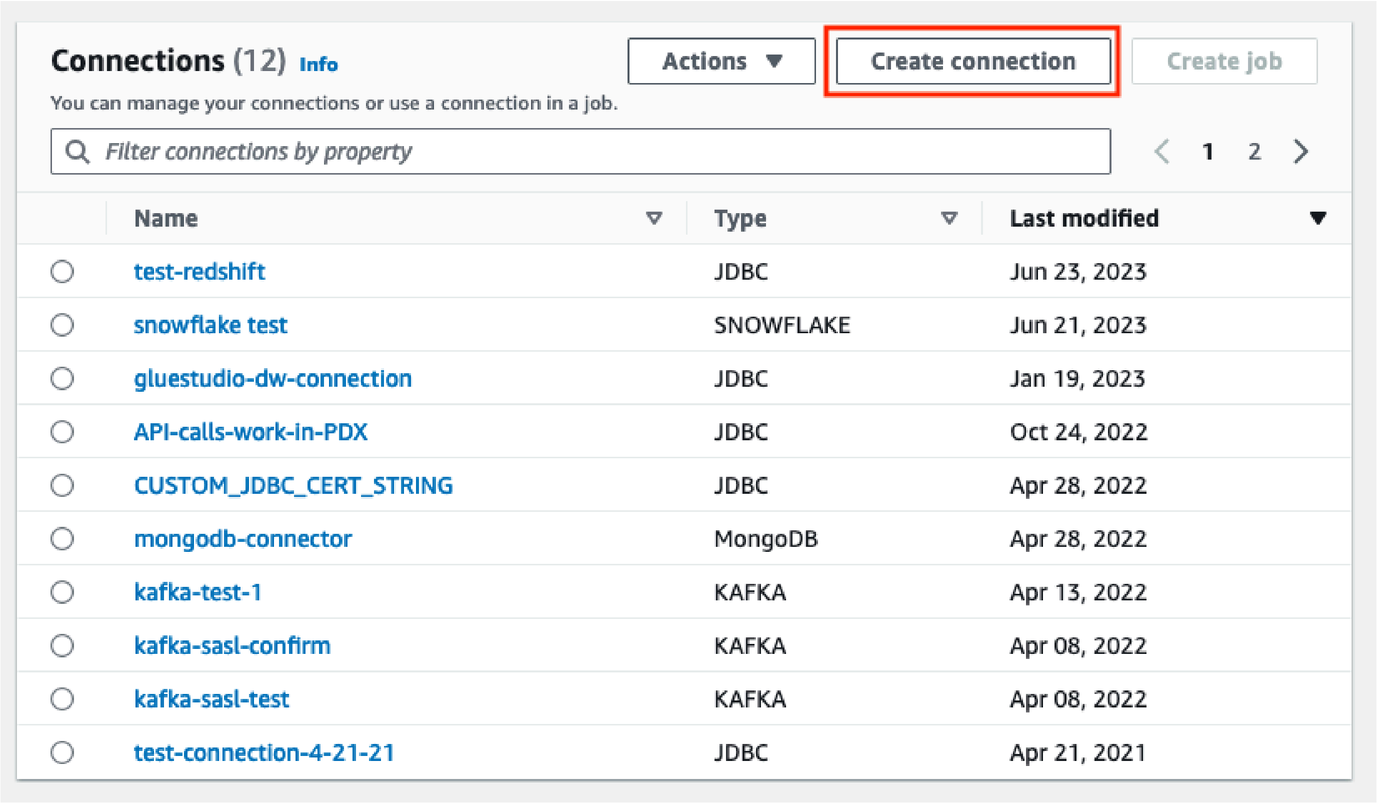
An AWS Glue connection is a Data Catalog object that stores connection information for a particular data store. Connections store login credentials, URI strings, virtual private cloud (VPC) information, and more. Creating connections in the Data Catalog saves the effort of having to specify all connection details every time you create a job.

To create a connection for a connector

1. In the AWS Glue Studio console, choose **Connectors** in the console navigation pane. In the **Connections** section, choose **Create connection**.
2. Choose the data source you want to create a connection for in step 1 of the **Create data connection** wizard. There are several ways to view the available data sources, including:
   - Filter the available data sources by choosing a tab. By default, **All connectors** is selected.
   - Toggle **List** to view the data sources as a list or toggle back to **Grid** to view the available connectors in the grid layout.
   - Use the search bar to narrow the list of data sources. As you type, search matches are displayed and non-matching sources are removed from view.

   Once you've chosen the data source, choose **Next**.
3. Configure the connection in Step 2 in the wizard.

   Enter the connection details. Depending on the type of connector you selected, you're prompted to enter additional information:



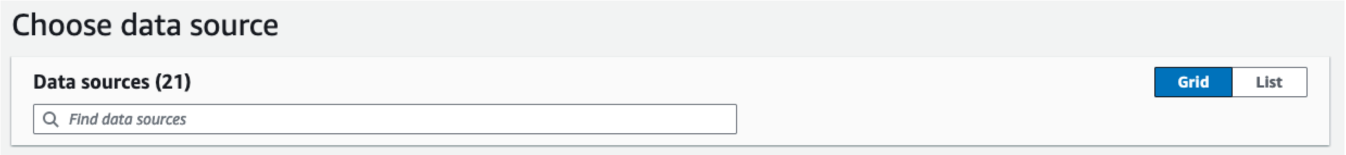
Connections (12)

You can manage your connections or use a connection in a job.

	Name	Type	Last modified
○	test-redshift	JDBC	Jun 23, 2023
○	snowflake test	SNOWFLAKE	Jun 21, 2023
○	gluestudio-dw-connection	JDBC	Jan 19, 2023
○	API-calls-work-in-PDX	JDBC	Oct 24, 2022
○	CUSTOM_JDBC_CERT_STRING	JDBC	Apr 28, 2022
○	mongodb-connector	MongoDB	Apr 28, 2022
○	kafka-test-1	KAFKA	Apr 13, 2022
○	kafka-sasl-confirm	KAFKA	Apr 08, 2022
○	kafka-sasl-test	KAFKA	Apr 08, 2022
○	test-connection-4-21-21	JDBC	Apr 21, 2021

4. Choose the data source you want to create a connection for in step 1 of the **Create data connection** wizard. There are several ways to view the available data sources. By default, you will see all available data sources in a grid layout. You can also:

   - Toggle **List** to view the data sources as a list or toggle back to **Grid** to view the available connectors in the grid layout.
   - Use the search bar to narrow the list of data sources. As you type, search matches are displayed and non-matching sources are removed from view.



Choose data source

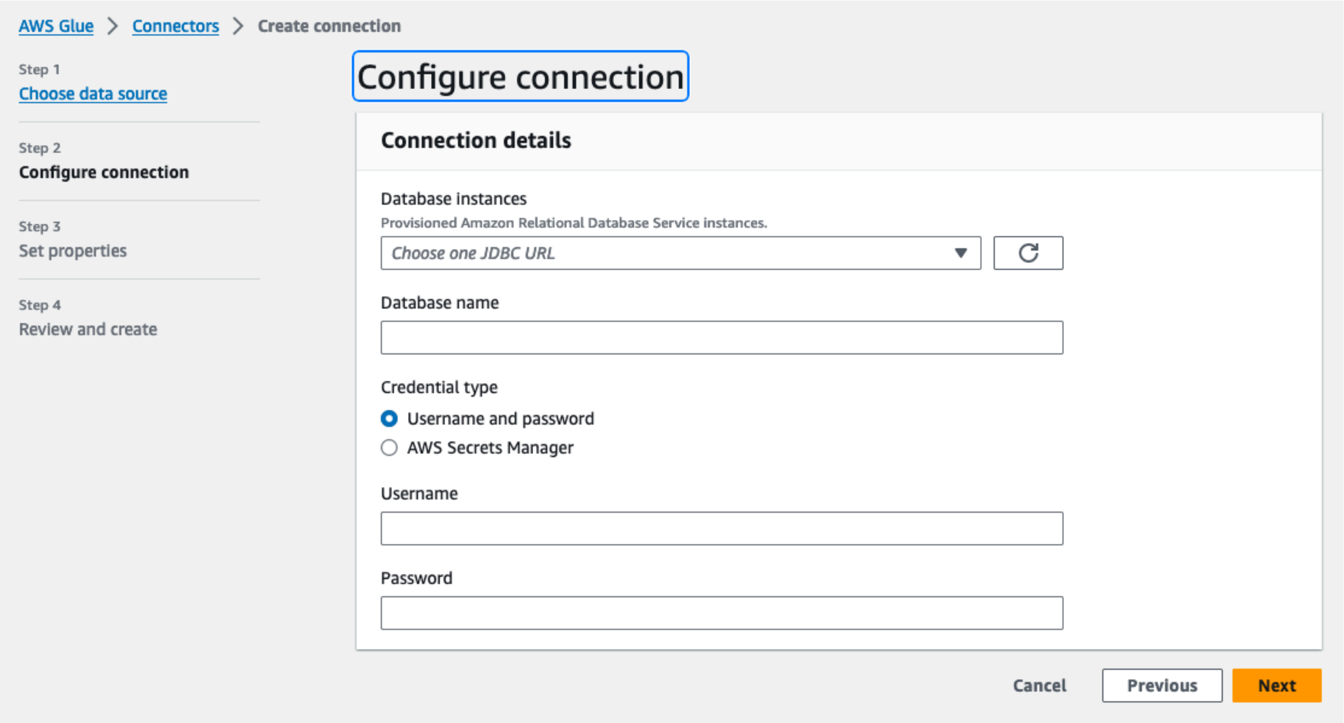
Data sources (21)

   Once you've chosen the data source, choose **Next**.

5. Configure the connection in Step 2 in the wizard.

   Enter the connection details. Depending on the type of connector you selected, you may be required to enter additional connection information. This can include:

- **Connection details** – these fields will change depending on the data source you are connecting to. For example, if you are connecting to Amazon DocumentDB databases, you will enter the Amazon DocumentDB URL. If you are connecting to Amazon Aurora, you will choose the database instance and enter the database name. The following is the Connection details required for Amazon Aurora:


AWS Glue > Connectors > Create connection
Step 1
Choose data source
Step 2
Configure connection
Step 3
Set properties
Step 4
Review and create
Configure connection
Connection details
Database instances
Provisioned Amazon Relational Database Service instances.
Choose one JDBC URL
Database name
Credential type
● Username and password
○ AWS Secrets Manager
Username
Password
Cancel
Previous
Next

- Credential type – choose between **Username and password** or **AWS Secrets Manager**. Enter the requested authentication information.
- For connectors that use JDBC, enter the information required to create the JDBC URL for the data store.
- If you use a virtual private cloud (VPC), then enter the network information for your VPC.

6. Set the connection properties in step 3 of the wizard. You can add a description and tags as an optional part of this step. Name is required and is prepopulated with a default value. Choose **Next**.
7. Review the connection source, details, and properties. If you need to make any changes, choose **Edit** for the step in the wizard. When ready, choose, **Create connection** .

   Choose **Create connection**.

   You are returned to the **Connectors** page, and the informational banner indicates the connection that was created. You can now use the connection in your AWS Glue Studio jobs.

Creating a Kafka connection

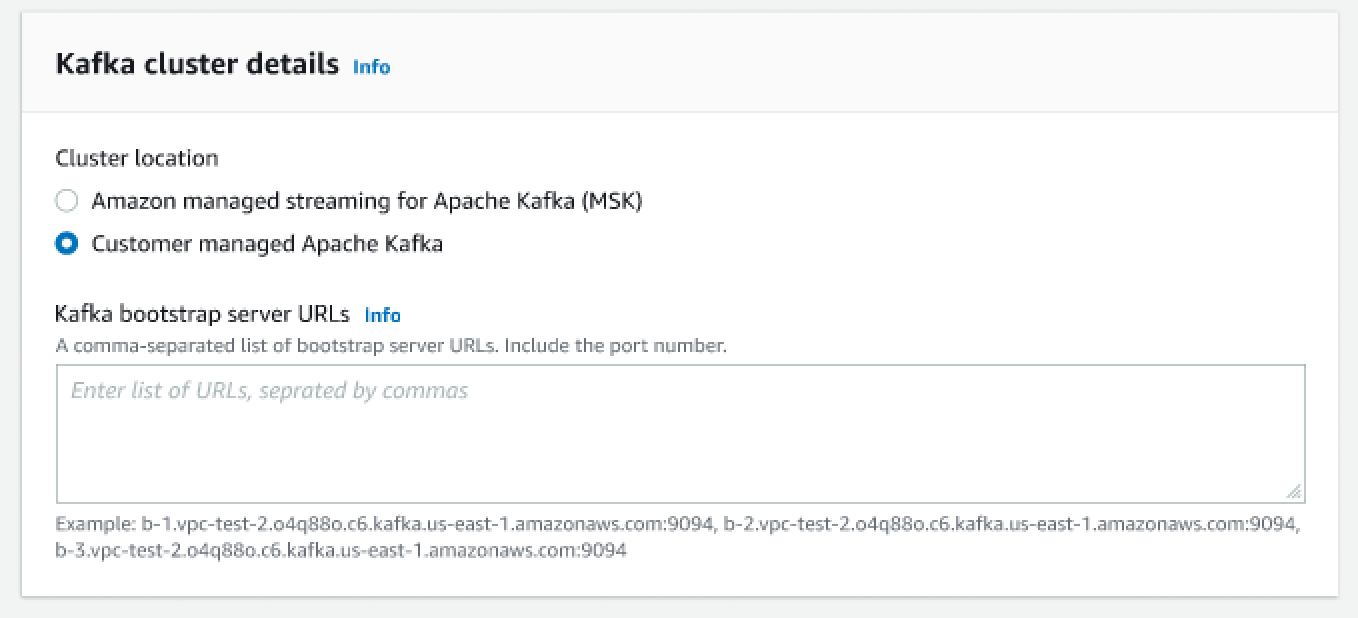
When creating a Kafka connection, selecting **Kafka** from the drop-down menu will display additional settings to configure:

- Kafka cluster details
- Authentication
- Encryption
- Network options

Configure Kafka cluster details

1. Choose the cluster location. You can choose from an **Amazon managed streaming for Apache Kafka (MSK)** cluster or a **Customer managed Apache Kafka** cluster. For more information on Amazon Managed streaming for Apache Kafka, see [Amazon managed streaming for Apache Kafka (MSK)]().

   > **Note**
   >
   > Amazon Managed Streaming for Apache Kafka only supports TLS and SASL/SCRAM-SHA-512 authentication methods.



Kafka cluster details

Cluster location

○ Amazon managed streaming for Apache Kafka (MSK)

● Customer managed Apache Kafka

Kafka bootstrap server URLs

A comma-separated list of bootstrap server URLs. Include the port number.

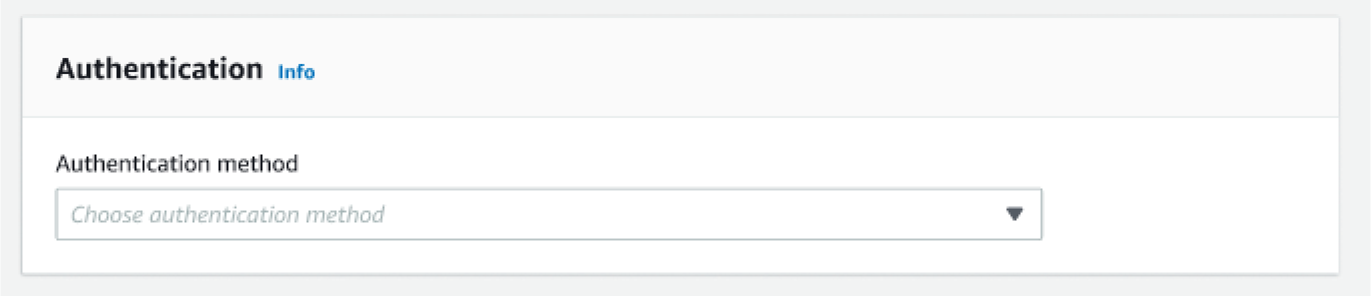
Enter list of URLs, seprated by commas

Example: b-1.vpc-test-2.o4q88o.c6.kafka.us-east-1.amazonaws.com:9094, b-2.vpc-test-2.o4q88o.c6.kafka.us-east-1.amazonaws.com:9094, b-3.vpc-test-2.o4q88o.c6.kafka.us-east-1.amazonaws.com:9094

2. Enter the URLs for your Kafka bootstrap servers. You may enter more than one by separating each server by a comma. Include the port number at the end of the URL by appending `:<port number>`.

   For example: `b-1.vpc-test-2.034a88o.kafka-us-east-1.amazonaws.com:9094`

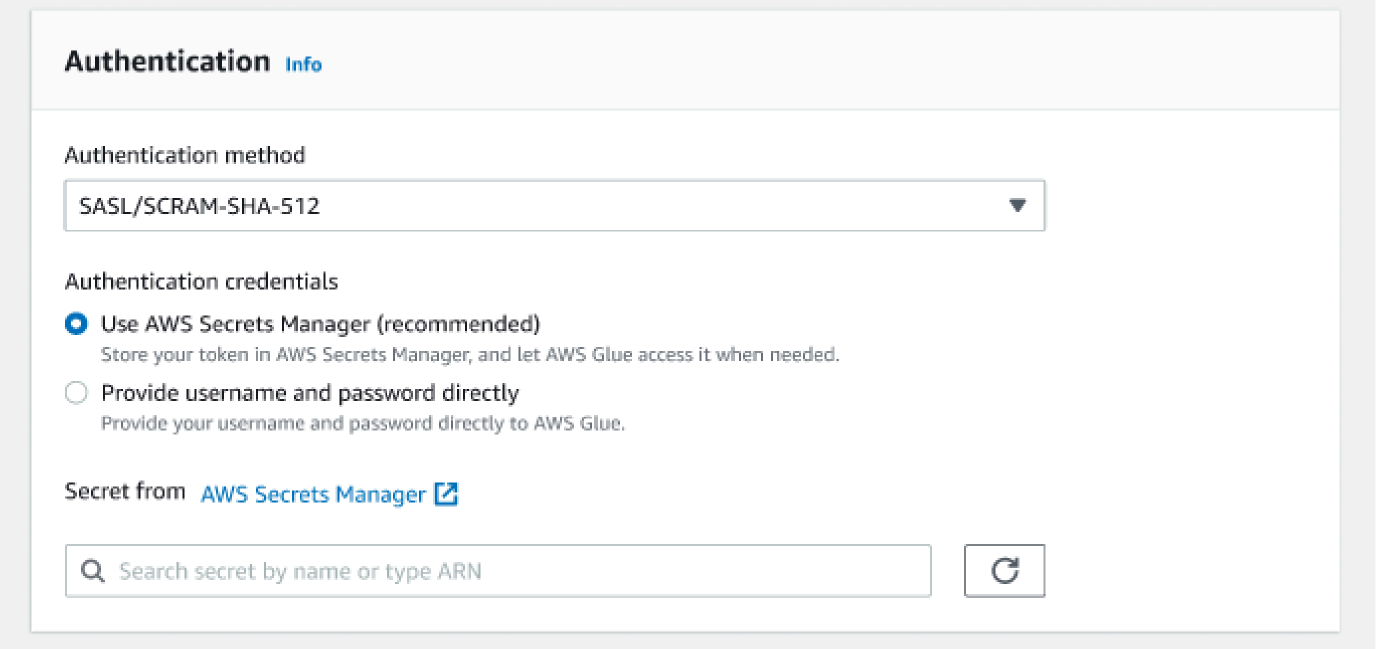
Select authentication method



AWS Glue supports the Simple Authentication and Security Layer (SASL) framework for authentication. The SASL framework supports various mechanisms of authentication, and AWS Glue offers the SCRAM (username and password), GSSAPI (Kerberos protocol), and PLAIN (username and password) protocols.

When choosing an authentication method from the drop-down menu, the following client authentication methods can be selected:

- None - No authentication. This is useful if you create a connection for testing purposes.
- SASL/SCRAM-SHA-512 - Choose this authentication method to specify authentication credentials. There are two options available:
  - Use AWS Secrets Manager (recommended) - if you select this option, you can store your credentials in AWS Secrets Manager and let AWS Glue access the information when needed. Specify the secret that stores the SSL or SASL authentication credentials.



Authentication Info

Authentication method

SASL/SCRAM-SHA-512

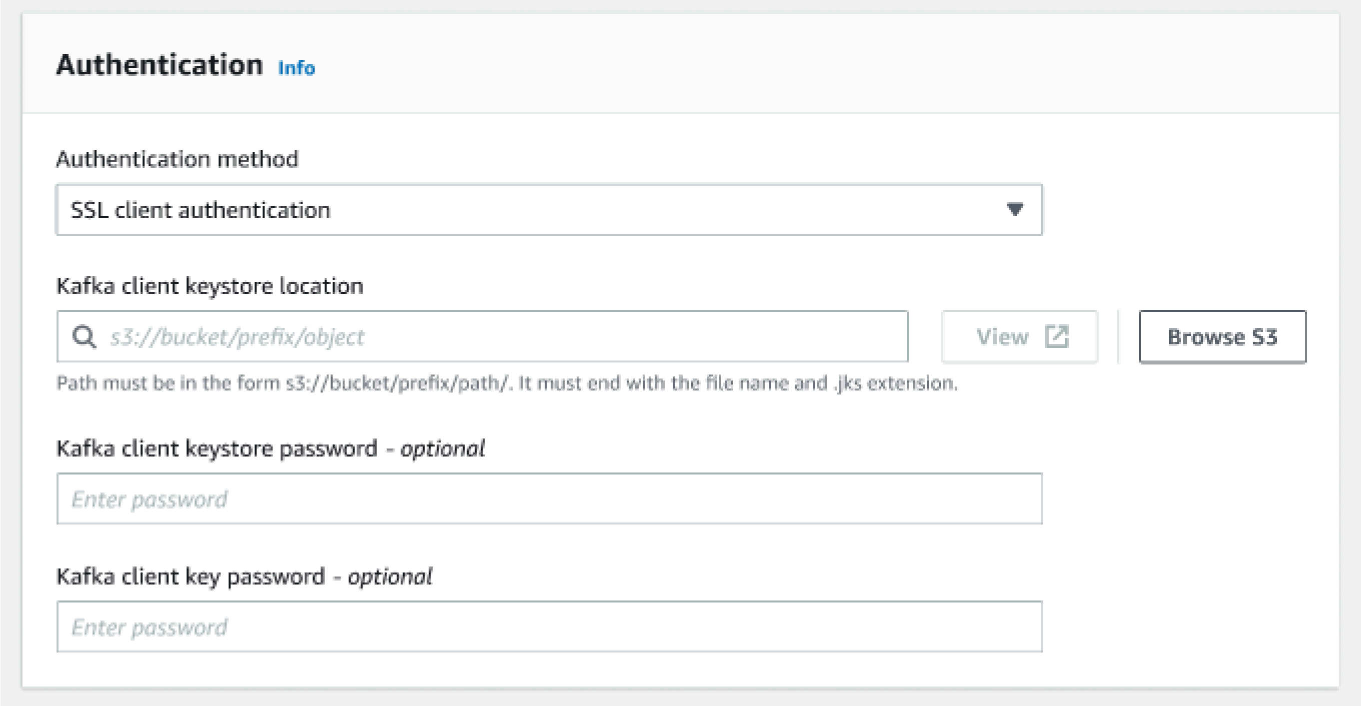
Authentication credentials

● Use AWS Secrets Manager (recommended)
Store your token in AWS Secrets Manager, and let AWS Glue access it when needed.

○ Provide username and password directly
Provide your username and password directly to AWS Glue.

Secret from AWS Secrets Manager

  - Provide username and password directly.
- SASL/GSSAPI (Kerberos) - if you select this option, you can select the location of the keytab file, krb5.conf file and enter the Kerberos principal name and Kerberos service name. The locations for the keytab file and krb5.conf file must be in an Amazon S3 location. Since MSK does not yet support SASL/GSSAPI, this option is only available for customer managed Apache Kafka clusters. For more information, see MIT Kerberos Documentation: Keytab .
- SASL/PLAIN - Choose this authentication method to specify authentication credentials. There are two options available:
  - Use AWS Secrets Manager (recommended) - if you select this option, you can store your credentials in AWS Secrets Manager and let AWS Glue access the information when needed. Specify the secret that stores the SSL or SASL authentication credentials.
  - Provide username and password directly.
- SSL Client Authentication - if you select this option, you can you can select the location of the Kafka client keystore by browsing Amazon S3. Optionally, you can enter the Kafka client keystore password and Kafka client key password.



Authentication Info

Authentication method

SSL client authentication ▼

Kafka client keystore location

s3://bucket/prefix/object

Path must be in the form s3://bucket/prefix/path/. It must end with the file name and .jks extension.

Kafka client keystore password - *optional*

Enter password

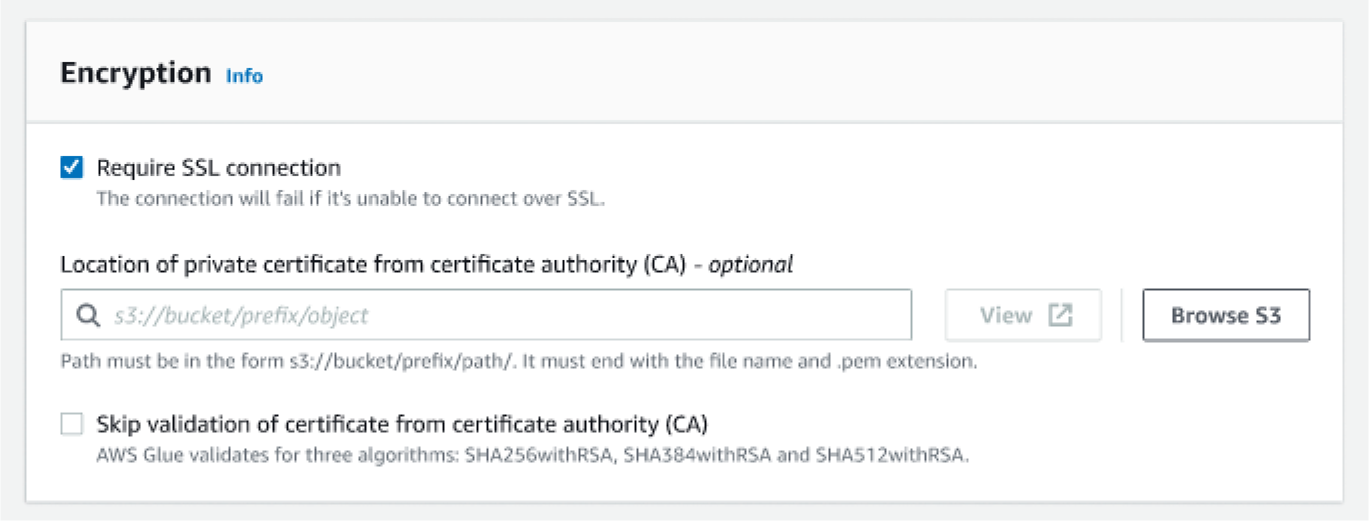
Kafka client key password - *optional*

Enter password

Configure encryption settings

1. If the Kafka connection requires SSL connection, select the checkbox for **Require SSL connection**. Note that the connection will fail if it's unable to connect over SSL. SSL for encryption can be used with any of the authentication methods (SASL/SCRAM-SHA-512, SASL/GSSAPI, SASL/PLAIN, or SSL Client Authentication) and is optional.

   If the authentication method is set to **SSL client authentication**, this option will be selected automatically and will be disabled to prevent any changes.
2. (Optional). Choose the location of private certificate from certificate authority (CA). Note that the location of the certification must be in an S3 location. Choose **Browse** to choose the file from a connected S3 bucket. The path must be in the form `s3://bucket/prefix/filename.pem`. It must end with the file name and .pem extension.
3. You can choose to skip validation of certificate from a certificate authority (CA). Choose the checkbox **Skip validation of certificate from certificate authority (CA)**. If this box is not checked, AWS Glue validates certificates for three algorithms:
   - SHA256withRSA
   - SHA384withRSA
   - SHA512withRSA



Encryption

☑ Require SSL connection
The connection will fail if it's unable to connect over SSL.

Location of private certificate from certificate authority (CA) - *optional*

s3://bucket/prefix/object

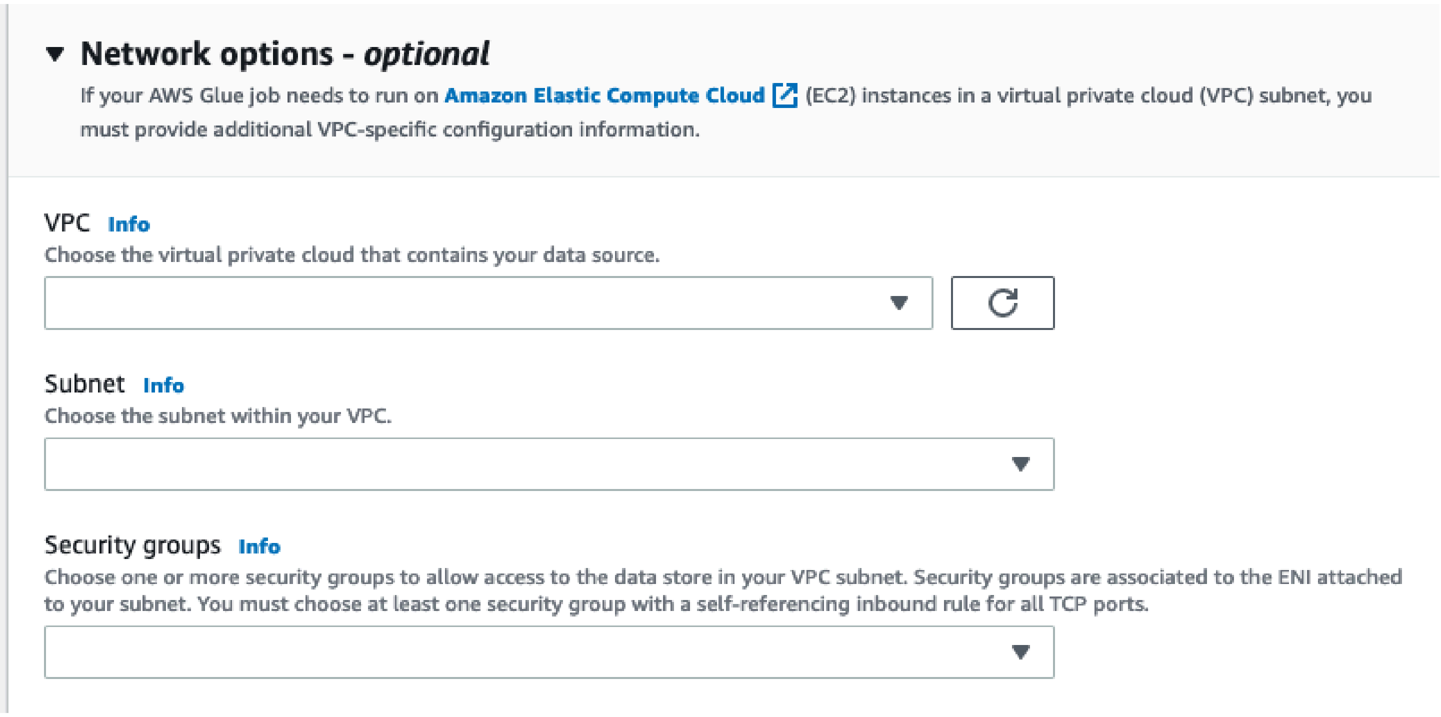
Path must be in the form s3://bucket/prefix/path/. It must end with the file name and .pem extension.

☐ Skip validation of certificate from certificate authority (CA)
AWS Glue validates for three algorithms: SHA256withRSA, SHA384withRSA and SHA512withRSA.

(Optional) Network options

The following are optional steps to configure VPC, Subnet and Security groups. If your AWS Glue job needs to run on Amazon EC2 instances in a virtual private cloud (VPC) subnet, you must provide additional VPC-specific configuration information.

1. Choose the VPC (virtual private cloud) that contains your data source.
2. Choose the subnet with your VPC.
3. Choose one or more security groups to allow access to the data store in your VPC subnet. Security groups are associated to the ENI attached to your subnet. You must choose at least one security group with a self-referencing inbound rule for all TCP ports.



▼ Network options - *optional*

If your AWS Glue job needs to run on **Amazon Elastic Compute Cloud** (EC2) instances in a virtual private cloud (VPC) subnet, you must provide additional VPC-specific configuration information.

VPC Info

Choose the virtual private cloud that contains your data source.

Subnet Info

Choose the subnet within your VPC.

Security groups Info

Choose one or more security groups to allow access to the data store in your VPC subnet. Security groups are associated to the ENI attached to your subnet. You must choose at least one security group with a self-referencing inbound rule for all TCP ports.

Authoring jobs with custom connectors

You can use connectors and connections for both data source nodes and data target nodes in AWS Glue Studio.

Topics

- Create jobs that use a connector for the data source
- Configure source properties for nodes that use connectors
- Configure target properties for nodes that use connectors

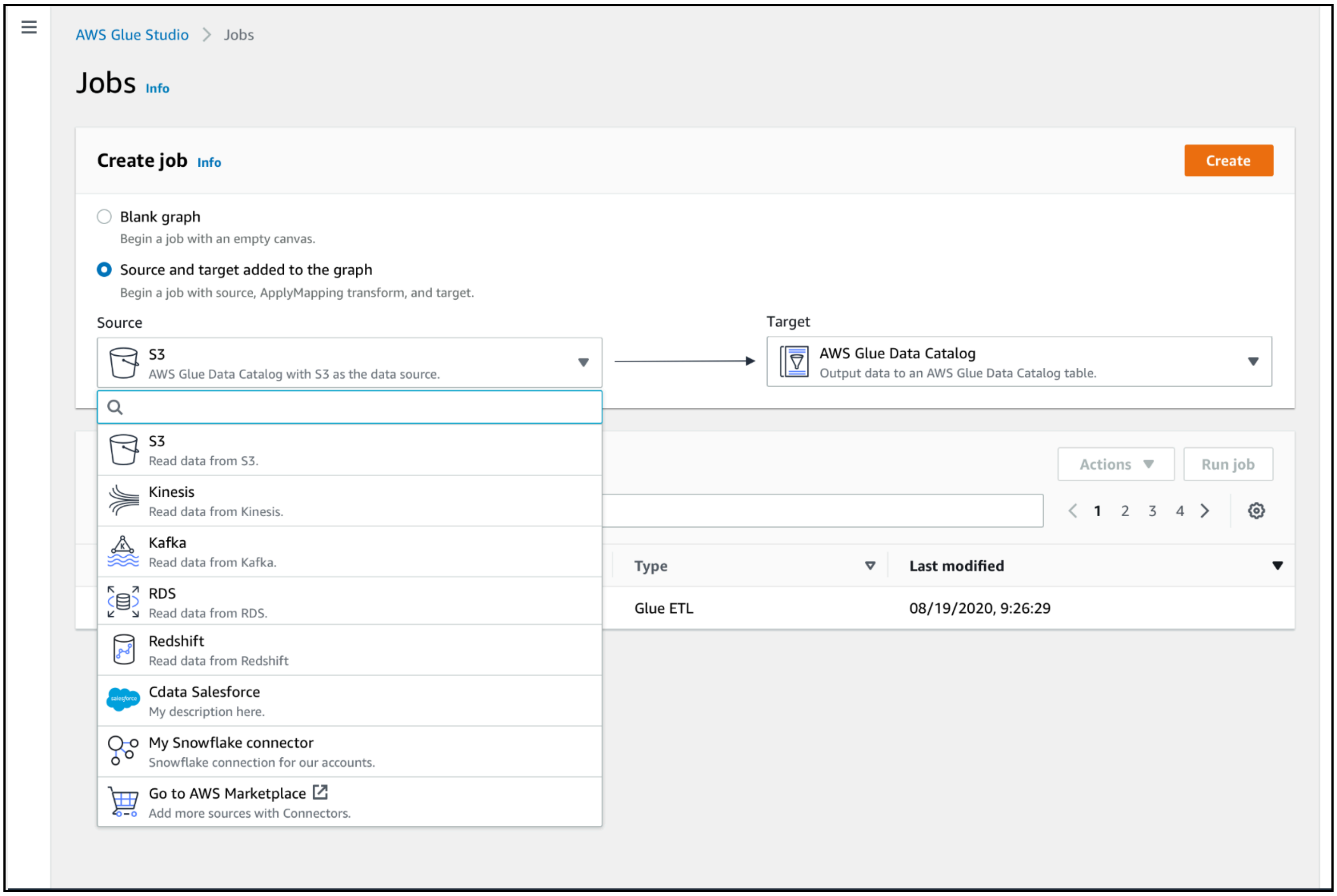
Create jobs that use a connector for the data source

When you create a new job, you can choose a connector for the data source and data targets.

To create a job that uses connectors for the data source or data target

1. Sign in to the AWS Management Console and open the AWS Glue Studio console at https://console.aws.amazon.com/gluestudio/.
2. On the **Connectors** page, in the **Your connections** resource list, choose the connection you want to use in your job, and then choose **Create job**.

Alternatively, on the AWS Glue Studio **Jobs** page, under **Create job**, choose **Source and target added to the graph**. In the **Source** drop-down list, choose the custom connector that you want to use in your job. You can also choose a connector for **Target**.


AWS Glue Studio > Jobs
Jobs Info
Create job Info
Create
Blank graph
Begin a job with an empty canvas.
Source and target added to the graph
Begin a job with source, ApplyMapping transform, and target.
Source
S3
AWS Glue Data Catalog with S3 as the data source.
Target
AWS Glue Data Catalog
Output data to an AWS Glue Data Catalog table.
S3
Read data from S3.
Kinesis
Read data from Kinesis.
Kafka
Read data from Kafka.
RDS
Read data from RDS.
Redshift
Read data from Redshift
Cdata Salesforce
My description here.
My Snowflake connector
Snowflake connection for our accounts.
Go to AWS Marketplace
Add more sources with Connectors.
Actions
Run job
1 2 3 4
Type
Last modified
Glue ETL
08/19/2020, 9:26:29

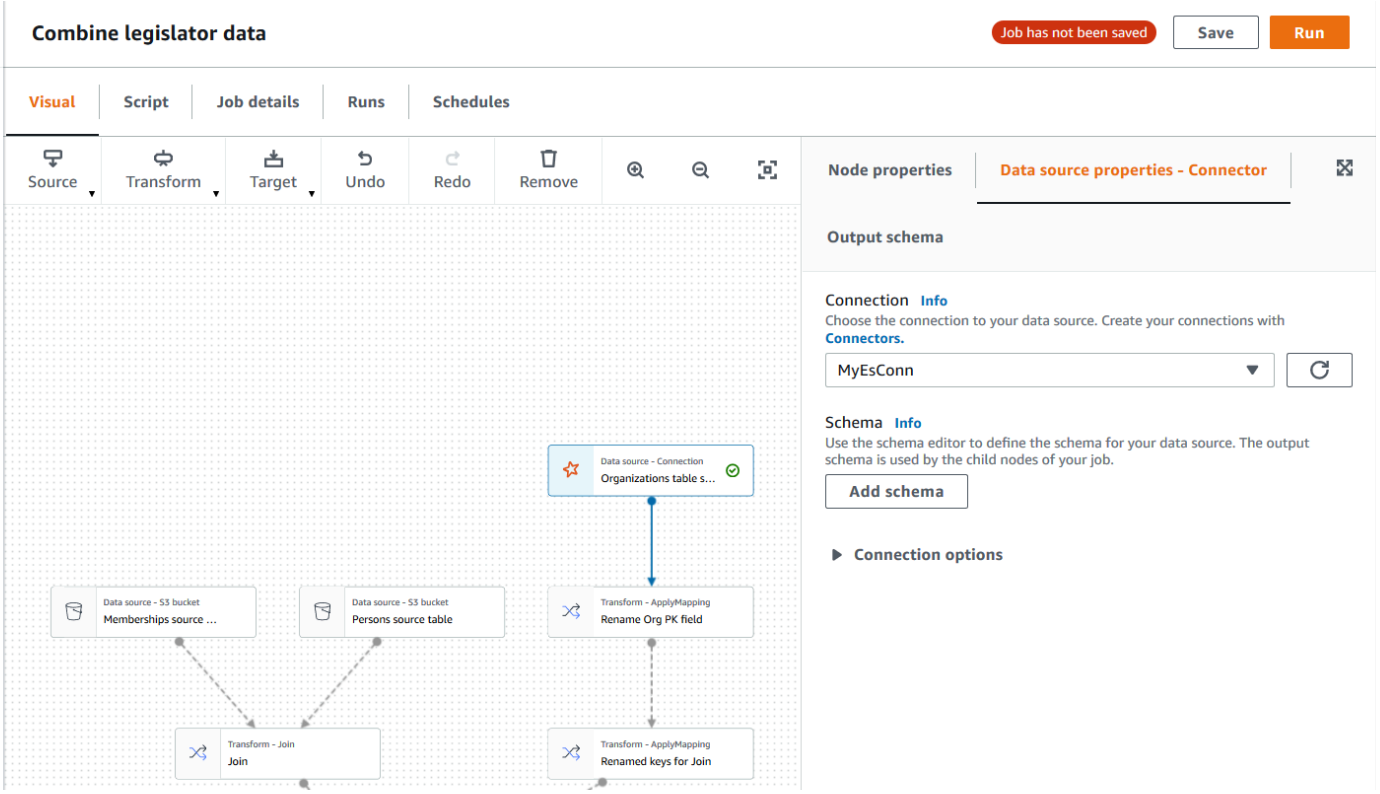
3. Choose **Create** to open the visual job editor.
4. Configure the data source node, as described in [Configure source properties for nodes that use connectors](#).
5. Continue creating your ETL job by adding transforms, additional data stores, and data targets, as described in [Starting visual ETL jobs in AWS Glue Studio](#).
6. Customize the job run environment by configuring job properties as described in [Modify the job properties](#).
7. Save and run the job.

Configure source properties for nodes that use connectors

After you create a job that uses a connector for the data source, the visual job editor displays a job graph with a data source node configured for the connector. You must configure the data source properties for that node.

To configure the properties for a data source node that uses a connector

1. Choose the connector data source node in the job graph or add a new node and choose the connector for the **Node type**. Then, on the right-side, in the node details panel, choose the **Data source properties** tab, if it's not already selected.


Combine legislator data
Job has not been saved
Save
Run
Visual
Script
Job details
Runs
Schedules
Source
Transform
Target
Undo
Redo
Remove
Node properties
Data source properties - Connector
Output schema
Data source - Connection
Organizations table s...
Data source - S3 bucket
Memberships source ...
Data source - S3 bucket
Persons source table
Transform - ApplyMapping
Rename Org PK field
Transform - Join
Join
Transform - ApplyMapping
Renamed keys for Join
Add schema
Connection options

Connection Info

Choose the connection to your data source. Create your connections with Connectors.

MyEsConn

Schema Info

Use the schema editor to define the schema for your data source. The output schema is used by the child nodes of your job.

2. In the **Data source properties** tab, choose the connection that you want to use for this job.

   Enter the additional information required for each connection type:

   JDBC

   - **Data source input type**: Choose to provide either a table name or a SQL query as the data source. Depending on your choice, you then need to provide the following additional information:

  - **Table name**: The name of the table in the data source. If the data source does not use the term *table*, then supply the name of an appropriate data structure, as indicated by the custom connector usage information (which is available in AWS Marketplace).
  - **Filter predicate**: A condition clause to use when reading the data source, similar to a `WHERE` clause, which is used to retrieve a subset of the data.
  - **Query code**: Enter a SQL query to use to retrieve a specific dataset from the data source. An example of a basic SQL query is:

    ```
    SELECT column_list FROM
                                table_name WHERE where_clause
    ```

- **Schema**: Because AWS Glue Studio is using information stored in the connection to access the data source instead of retrieving metadata information from a Data Catalog table, you must provide the schema metadata for the data source. Choose **Add schema** to open the schema editor.

  For instructions on how to use the schema editor, see Editing the schema in a custom transform node.

- **Partition column**: (Optional) You can choose to partition the data reads by providing values for **Partition column**, **Lower bound**, **Upper bound**, and **Number of partitions**.

  The `lowerBound` and `upperBound` values are used to decide the partition stride, not for filtering the rows in table. All rows in the table are partitioned and returned.

  > **Note**
  >
  > Column partitioning adds an extra partitioning condition to the query used to read the data. When using a query instead of a table name, you should validate that the query works with the specified partitioning condition. For example:
  >
  > - If your query format is `"SELECT col1 FROM table1"`, then test the query by appending a `WHERE` clause at the end of the query that uses the partition column.
  > - If your query format is `"SELECT col1 FROM table1 WHERE col2=val"`, then test the query by extending the `WHERE` clause with `AND` and an expression that uses the partition column.

- **Data type casting**: If the data source uses data types that are not available in JDBC, use this section to specify how a data type from the data source should be converted into JDBC data types. You can specify up to 50 different data type conversions. All columns in the data source that use the same data type are converted in the same way.

  For example, if you have three columns in the data source that use the `Float` data type, and you indicate that the `Float` data type should be converted to the JDBC `String` data type, then all three columns that use the `Float` data type are converted to `String` data types.
- **Job bookmark keys**: Job bookmarks help AWS Glue maintain state information and prevent the reprocessing of old data. Specify one more one or more columns as bookmark keys. AWS Glue Studio uses bookmark keys to track data that has already been processed during a previous run of the ETL job. Any columns you use for custom bookmark keys must be strictly monotonically increasing or decreasing, but gaps are permitted.

  If you enter multiple bookmark keys, they're combined to form a single compound key. A compound job bookmark key should not contain duplicate columns. If you don't specify bookmark keys, AWS Glue Studio by default uses the primary key as the bookmark key, provided that the primary key is sequentially increasing or decreasing (with no gaps). If the table doesn't have a primary key, but the job bookmark property is enabled, you must provide custom job bookmark keys. Otherwise, the search for primary keys to use as the default will fail and the job run will fail.
- **Job bookmark keys sorting order**: Choose whether the key values are sequentially increasing or decreasing.

Spark

- **Schema**: Because AWS Glue Studio is using information stored in the connection to access the data source instead of retrieving metadata information from a Data Catalog table, you must provide the schema metadata for the data source. Choose **Add schema** to open the schema editor.

  For instructions on how to use the schema editor, see Editing the schema in a custom transform node.

- **Connection options**: Enter additional key-value pairs as needed to provide additional connection information or options. For example, you might enter a database name, table name, a user name, and password.

  For example, for OpenSearch, you enter the following key-value pairs, as described in the section called " Tutorial: Using the AWS Glue Connector for Elasticsearch ":

  - `es.net.http.auth.user` : *`username`*
  - `es.net.http.auth.pass` : *`password`*
  - `es.nodes` : `https://`*`<Elasticsearch endpoint>`*
  - `es.port` : `443`
  - `path`: *`<Elasticsearch resource>`*
  - `es.nodes.wan.only` : `true`

For an example of the minimum connection options to use, see the sample test script MinimalSparkConnectorTest.scala on GitHub, which shows the connection options you would normally provide in a connection.

Athena

- **Table name**: The name of the table in the data source. If you're using a connector for reading from Athena-CloudWatch logs, you would enter the table name `all_log_streams`.
- **Athena schema name**: Choose the schema in your Athena data source that corresponds to the database that contains the table. If you're using a connector for reading from Athena-CloudWatch logs, you would enter a schema name similar to `/aws/glue/`*`name`*.
- **Schema**: Because AWS Glue Studio is using information stored in the connection to access the data source instead of retrieving metadata information from a Data Catalog table, you must provide the schema metadata for the data source. Choose **Add schema** to open the schema editor.

  For instructions on how to use the schema editor, see Editing the schema in a custom transform node.
- **Additional connection options**: Enter additional key-value pairs as needed to provide additional connection information or options.

For an example, see the `README.md` file at https://github.com/aws-samples/aws-glue-samples/tree/master/GlueCustomConnectors/development/Athena. In the steps in this document, the sample code shows the minimal required connection options, which are `tableName`, `schemaName`, and `className`. The code example specifies these options as part of the `optionsMap` variable, but you can specify them for your connection and then use the connection.

3. (Optional) After providing the required information, you can view the resulting data schema for your data source by choosing the **Output schema** tab in the node details panel. The schema displayed on this tab is used by any child nodes that you add to the job graph.
4. (Optional) After configuring the node properties and data source properties, you can preview the dataset from your data source by choosing the Data preview tab in the node details panel. The first time you choose this tab for any node in your job, you are prompted to provide an IAM role to access the data. There is a cost associated with using this feature, and billing starts as soon as you provide an IAM role.

Configure target properties for nodes that use connectors

If you use a connector for the data target type, you must configure the properties of the data target node.

To configure the properties for a data target node that uses a connector

1. Choose the connector data target node in the job graph. Then, on the right-side, in the node details panel, choose the **Data target properties** tab, if it's not already selected.
2. In the **Data target properties** tab, choose the connection to use for writing to the target.

   Enter the additional information required for each connection type:

   JDBC

   - **Connection**: Choose the connection to use with your connector. For information about how to create a connection, see [Creating connections for connectors](#).
   - **Table name**: The name of the table in the data target. If the data target does not use the term *table*, then supply the name of an appropriate data structure, as indicated by the custom connector usage information (which is available in AWS Marketplace).

- **Batch size** (Optional): Enter the number of rows or records to insert in the target table in a single operation. The default value is 1000 rows.

Spark

- **Connection**: Choose the connection to use with your connector. If you did not create a connection previously, choose **Create connection** to create one. For information about how to create a connection, see Creating connections for connectors.
- **Connection options**: Enter additional key-value pairs as needed to provide additional connection information or options. You might enter a database name, table name, a user name, and password.

  For example, for OpenSearch, you enter the following key-value pairs, as described in the section called " Tutorial: Using the AWS Glue Connector for Elasticsearch ":

  - `es.net.http.auth.user`:*`username`*
  - `es.net.http.auth.pass`:*`password`*
  - `es.nodes`:`https://`*`<Elasticsearch endpoint>`*
  - `es.port`:`443`
  - `path`: *`<Elasticsearch resource>`*
  - `es.nodes.wan.only`:`true`

  For an example of the minimum connection options to use, see the sample test script MinimalSparkConnectorTest.scala on GitHub, which shows the connection options you would normally provide in a connection.

3. After providing the required information, you can view the resulting data schema for your data source by choosing the **Output schema** tab in the node details panel.

Managing connectors and connections

You use the **Connections** page in AWS Glue to manage your connectors and connections.

Topics

- Viewing connector and connection details
- Editing connectors and connections

- Deleting connectors and connections
- Cancel a subscription for a connector

Viewing connector and connection details

You can view summary information about your connectors and connections in the **Your connectors** and **Your connections** resource tables on the **Connectors** page. To view detailed information, perform the following steps.

To view connector or connection details

1. In the AWS Glue Studio console, choose **Connectors** in the console navigation pane.
2. Choose the connector or connection that you want to view detailed information for.
3. Choose **Actions**, and then choose **View details** to open the detail page for that connector or connection.
4. On the detail page, you can choose to **Edit** or **Delete** the connector or connection.

    - For connectors, you can choose **Create connection** to create a new connection that uses the connector.
    - For connections, you can choose **Create job** to create a job that uses the connection.

Editing connectors and connections

You use the **Connectors** page to change the information stored in your connectors and connections.

To modify a connector or connection

1. In the AWS Glue Studio console, choose **Connectors** in the console navigation pane.
2. Choose the connector or connection that you want to change.
3. Choose **Actions**, and then choose **Edit**.

    You can also choose **View details** and on the connector or connection detail page, you can choose **Edit**.
4. On the **Edit connector** or **Edit connection** page, update the information, and then choose **Save**.

Deleting connectors and connections

You use the **Connectors** page to delete connectors and connections. If you delete a connector, then any connections that were created for that connector should also be deleted.

To remove connectors from AWS Glue Studio

1. In the AWS Glue Studio console, choose **Connectors** in the console navigation pane.
2. Choose the connector or connection you want to delete.
3. Choose **Actions**, and then choose **Delete**.

   You can also choose **View details**, and on the connector or connection detail page, you can choose **Delete**.
4. Verify that you want to remove the connector or connection by entering **`Delete`**, and then choose **Delete**.

   When deleting a connector, any connections that were created for that connector are also deleted.

Any jobs that use a deleted connection will no longer work. You can either edit the jobs to use a different data store, or remove the jobs. For information about how to delete a job, see Delete jobs.

If you delete a connector, this doesn't cancel the subscription for the connector in AWS Marketplace. To remove a subscription for a deleted connector, follow the instructions in Cancel a subscription for a connector .

Cancel a subscription for a connector

After you delete the connections and connector from AWS Glue Studio, you can cancel your subscription in AWS Marketplace if you no longer need the connector.

> **Note**
>
> If you cancel your subscription to a connector, this does not remove the connector or connection from your account. Any jobs that use the connector and related connections will no longer be able to use the connector and will fail.
> Before you unsubscribe or re-subscribe to a connector from AWS Marketplace, you should delete existing connections and connectors associated with that AWS Marketplace product.

To unsubscribe from a connector in AWS Marketplace

1. Sign in to the AWS Marketplace console at https://console.aws.amazon.com/marketplace.
2. Choose **Manage subscriptions**.
3. On the **Manage subscriptions** page, choose **Manage** next to the connector subscription that you want to cancel.
4. Choose **Actions** and then choose **Cancel subscription**.
5. Select the check box to acknowledge that running instances are charged to your account, and then choose **Yes, cancel subscription**.

Developing custom connectors

You can write the code that reads data from or writes data to your data store and formats the data for use with AWS Glue Studio jobs. You can create connectors for Spark, Athena, and JDBC data stores. Sample code posted on GitHub provides an overview of the basic interfaces you need to implement.

You will need a local development environment for creating your connector code. You can use any IDE or even just a command line editor to write your connector. Examples of development environments include:

- A local Scala environment with a local AWS Glue ETL Maven library, as described in Developing Locally with Scala in the *AWS Glue Developer Guide*.
- IntelliJ IDE, by downloading the IDE from https://www.jetbrains.com/idea/.

Topics

- Developing Spark connectors
- Developing Athena connectors
- Developing JDBC connectors
- Examples of using custom connectors with AWS Glue Studio
- Developing AWS Glue connectors for AWS Marketplace

Developing Spark connectors

You can create a Spark connector with Spark DataSource API V2 (Spark 2.4) to read data.

To create a custom Spark connector

Follow the steps in the AWS Glue GitHub sample library for developing Spark connectors, which is located at https://github.com/aws-samples/aws-glue-samples/tree/master/GlueCustomConnectors/development/Spark/README.md.

Developing Athena connectors

You can create an Athena connector to be used by AWS Glue and AWS Glue Studio to query a custom data source.

To create a custom Athena connector

Follow the steps in the AWS Glue GitHub sample library for developing Athena connectors, which is located at https://github.com/aws-samples/aws-glue-samples/tree/master/GlueCustomConnectors/development/Athena.

Developing JDBC connectors

You can create a connector that uses JDBC to access your data stores.

To create a custom JDBC connector

1. Install the AWS Glue Spark runtime libraries in your local development environment. Refer to the instructions in the AWS Glue GitHub sample library at https://github.com/aws-samples/aws-glue-samples/tree/master/GlueCustomConnectors/development/GlueSparkRuntime/README.md.
2. Implement the JDBC driver that is responsible for retrieving the data from the data source. Refer to the Java Documentation for Java SE 8.

   Create an entry point within your code that AWS Glue Studio uses to locate your connector. The **Class name** field should be the full path of your JDBC driver.
3. Use the `GlueContext` API to read data with the connector. Users can add more input options in the AWS Glue Studio console to configure the connection to the data source, if necessary. For a code example that shows how to read from and write to a JDBC database with a custom JDBC connector, see Custom and AWS Marketplace connectionType values.

Examples of using custom connectors with AWS Glue Studio

You can refer to the following blogs for examples of using custom connectors:

- Developing, testing, and deploying custom connectors for your data stores with AWS Glue
- Apache Hudi: Writing to Apache Hudi tables using AWS Glue Custom Connector
- Google BigQuery: Migrating data from Google BigQuery to Amazon S3 using AWS Glue custom connectors
- Snowflake (JDBC): Performing data transformations using Snowflake and AWS Glue
- SingleStore: Building fast ETL using SingleStore and AWS Glue
- Salesforce: Ingest Salesforce data into Amazon S3 using the CData JDBC custom connector with AWS Glue -
- MongoDB: Building AWS Glue Spark ETL jobs using Amazon DocumentDB (with MongoDB compatibility) and MongoDB
- Amazon Relational Database Service (Amazon RDS): Building AWS Glue Spark ETL jobs by bringing your own JDBC drivers for Amazon RDS
- MySQL (JDBC): https://github.com/aws-samples/aws-glue-samples/blob/master/GlueCustomConnectors/development/Spark/SparkConnectorMySQL.scala

Developing AWS Glue connectors for AWS Marketplace

As an AWS partner, you can create custom connectors and upload them to AWS Marketplace to sell to AWS Glue customers.

The process for developing the connector code is the same as for custom connectors, but the process of uploading and verifying the connector code is more detailed. Refer to the instructions in Creating Connectors for AWS Marketplace on the GitHub website.

Restrictions for using connectors and connections in AWS Glue Studio

When you're using custom connectors or connectors from AWS Marketplace, take note of the following restrictions:

- The testConnection API isn't supported with connections created for custom connectors.
- Data Catalog connection password encryption isn't supported with custom connectors.
- You can't use job bookmarks if you specify a filter predicate for a data source node that uses a JDBC connector.
- Creating a Marketplace connection is not supported outside of the AWS Glue Studio user interface.

Testing an AWS Glue connection

As a best practice, before you use an AWS Glue connection in an ETL job, use the AWS Glue console to test the connection. AWS Glue uses the parameters in your connection to confirm that it can access your data store and reports any errors. For information about AWS Glue connections, see Connecting to data.

To test an AWS Glue connection

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, under **Data Catalog**, choose **Connections**. You can also choose **Data connections** above **Data Catalog** in the navigation pane.
3. In **Connections**, select the check box next to the desired connection, and then choose **Actions**. In the drop-down menu, choose **Test connection**.
4. In the **Test connection** dialog box, select a role or choose **Create IAM role** to go to the AWS Identity and Access Management (IAM) console to create a new role. The role must have permissions on the data store.
5. Choose **Confirm**.

   The test begins and can take several minutes to complete. If the test fails, choose **Troubleshoot** to view the steps to resolve the issue.
6. Choose **Logs** to view the logs in CloudWatch. You must have the required IAM permissions to view the logs. For more information, see AWS Managed (Predefined) Policies for CloudWatch Logs in the *Amazon CloudWatch Logs User Guide*.

Configuring AWS calls to go through your VPC

The special job parameter `disable-proxy-v2` allows you to route your calls to services such as Amazon S3, CloudWatch, and AWS Glue through your VPC. By default, AWS Glue uses a local proxy to send traffic through the AWS Glue VPC to download scripts and libraries from Amazon S3, to send requests to CloudWatch for publishing logs and metrics, and to send requests to AWS Glue for accessing data catalogs. This proxy allows the job to function normally even if your VPC doesn't configure a proper route to other AWS services, such as Amazon S3, CloudWatch, and AWS Glue. AWS Glue now offers a parameter for you to turn off this behavior. For more information, see Job parameters used by AWS Glue. AWS Glue will continue to use local proxy for publishing CloudWatch logs of your AWS Glue jobs.

> **Note**
>
> - This feature is supported for AWS Glue jobs with AWS Glue version 2.0 and above. When using this feature, you need to ensure that your VPC has configured a route to Amazon S3 through a NAT or service VPC endpoint.
> - The deprecated job parameter `disable-proxy` only routes your calls to Amazon S3 for downloading scripts and libraries through your VPC. It's recommended to use the new parameter `disable-proxy-v2` instead.

Example usage

Create an AWS Glue job with `disable-proxy-v2`:

```
aws glue create-job \
    --name no-proxy-job \
    --role GlueDefaultRole \
    --command "Name=glueetl,ScriptLocation=s3://my-bucket/glue-script.py" \
    --connections Connections="traffic-monitored-connection" \
    --default-arguments '{"--disable-proxy-v2" : "true"}'
```

Connecting to a JDBC data store in a VPC

Typically, you create resources inside Amazon Virtual Private Cloud (Amazon VPC) so that they cannot be accessed over the public internet. By default, AWS Glue can't access resources inside a VPC. To enable AWS Glue to access resources inside your VPC, you must provide additional VPC-specific configuration information that includes VPC subnet IDs and security group IDs. AWS Glue uses this information to set up elastic network interfaces that enable your function to connect securely to other resources in your private VPC.

When using a VPC endpoint, add it to your route table. For more information, see Creating an interface VPC endpoint for AWS Glue and Prerequisites.

When using encryption in Data Catalog, create the KMS interface endpoint and add it to your route table. For more information, see
Creating a VPC endpoint for AWS KMS.

Accessing VPC Data Using elastic network interfaces

When AWS Glue connects to a JDBC data store in a VPC, AWS Glue creates an elastic network interface (with the prefix `Glue_`) in your account to access your VPC data. You can't delete this network interface as long as it's attached to AWS Glue. As part of creating the elastic network interface, AWS Glue associates one or more security groups to it. To enable AWS Glue to create the network interface, security groups that are associated with the resource must allow inbound access with a source rule. This rule contains a security group that is associated with the resource. This gives the elastic network interface access to your data store with the same security group.

To allow AWS Glue to communicate with its components, specify a security group with a self-referencing inbound rule for all TCP ports. By creating a self-referencing rule, you can restrict the source to the same security group in the VPC and not open it to all networks. The default security group for your VPC might already have a self-referencing inbound rule for `ALL Traffic`.

You can create rules in the Amazon VPC console. To update rule settings via the AWS Management Console, navigate to the VPC console (https://console.aws.amazon.com/vpc/), and select the appropriate security group. Specify the inbound rule for `ALL TCP` to have as its source the same security group name. For more information about security group rules, see Security Groups for Your VPC.

Each elastic network interface is assigned a private IP address from the IP address range in the subnets that you specify. The network interface is not assigned any public IP addresses. AWS Glue requires internet access (for example, to access AWS services that don't have VPC endpoints). You can configure a network address translation (NAT) instance inside your VPC, or you can use the Amazon VPC NAT gateway. For more information, see NAT Gateways in the *Amazon VPC User Guide*. You can't directly use an internet gateway attached to your VPC as a route in your subnet route table because that requires the network interface to have public IP addresses.

The VPC network attributes `enableDnsHostnames` and `enableDnsSupport` must be set to true. For more information, see Using DNS with your VPC.

> ⚠ **Important**
>
> Don't put your data store in a public subnet or in a private subnet that doesn't have internet access. Instead, attach it only to private subnets that have internet access through a NAT instance or an Amazon VPC NAT gateway.

Elastic network interface properties

To create the elastic network interface, you must supply the following properties:

VPC

The name of the VPC that contains your data store.

Subnet

The subnet in the VPC that contains your data store.

Security groups

The security groups that are associated with your data store. AWS Glue associates these security groups with the elastic network interface that is attached to your VPC subnet. To allow AWS Glue components to communicate and also prevent access from other networks, at least one chosen security group must specify a self-referencing inbound rule for all TCP ports.

For information about managing a VPC with Amazon Redshift, see Managing Clusters in an Amazon Virtual Private Cloud (VPC).

For information about managing a VPC with Amazon Relational Database Service (Amazon RDS), see Working with an Amazon RDS DB Instance in a VPC.

Using a MongoDB or MongoDB Atlas connection

After you create a connection for MongoDB or MongoDB Atlas, you can use the connection in your ETL job. You create a table in the AWS Glue Data Catalog and specify the MongoDB or MongoDB Atlas connection for the `connection` attribute of the table.

AWS Glue stores your connection `url` and credentials in the MongoDB connection. The connection URI formats are as follows:

- For MongoDB: mongodb://host:port/database. The host can be a hostname, IP address, or UNIX domain socket. If the connection string doesn't specify a port, it uses the default MongoDB port, 27017.
- For MongoDB Atlas: mongodb+srv://server.example.com/database. The host can be a hostname that follows corresponds to a DNS SRV record. The SRV format does not require a port and will use the default MongoDB port, 27017.

Additionally, you can specify options in your job script. For more information, see the section called "MongoDB connection".

Crawling an Amazon S3 data store using a VPC endpoint

For security, auditing, or control purposes you may want your Amazon S3 data store or Amazon S3 backed Data Catalog tables to only be accessed through an Amazon Virtual Private Cloud environment (Amazon VPC). This topic describes how to create and test a connection to the Amazon S3 data store or Amazon S3 backed Data Catalog tables in a VPC endpoint using the `Network` connection type.

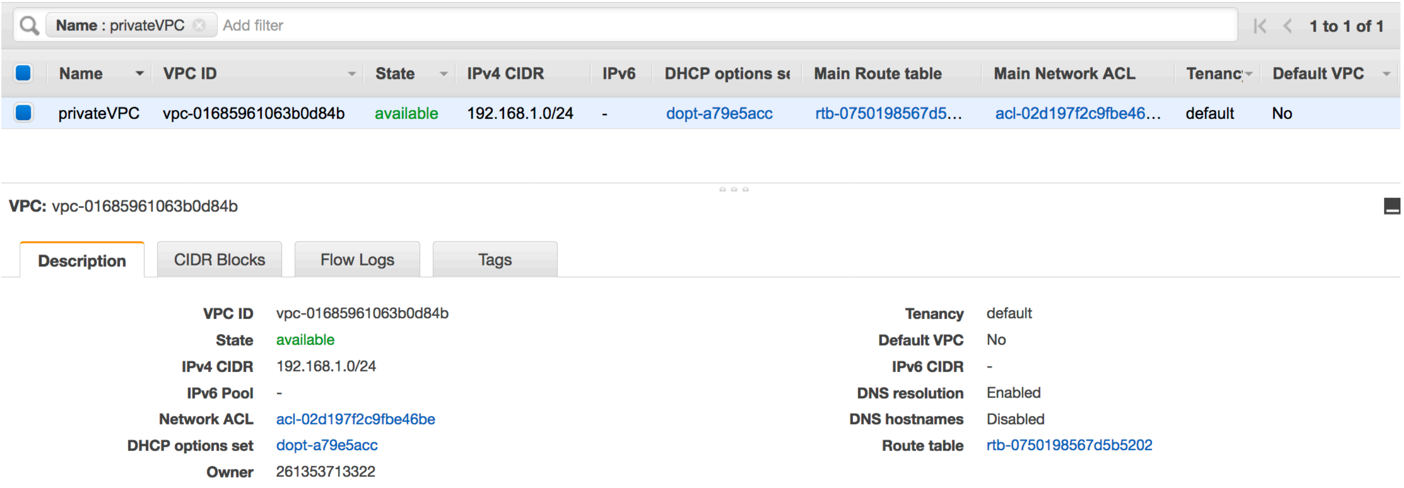
Perform the following tasks to run a crawler on the data store:

- the section called "Prerequisites"
- the section called "Creating the connection to Amazon S3"
- the section called "Testing the connection to Amazon S3"
- the section called "Creating a crawler for an Amazon S3 data store"
- the section called "Running a crawler"

Prerequisites

Check that you have met these prerequisites for setting up your Amazon S3 data store or Amazon S3 backed Data Catalog tables to be accessed through an Amazon Virtual Private Cloud environment (Amazon VPC).

- A configured VPC. For example: vpc-01685961063b0d84b. For more information, see Getting started with Amazon VPC in the *Amazon VPC User Guide*.
- An Amazon S3 endpoint attached to the VPC. For example: vpc-01685961063b0d84b. For more information, see Endpoints for Amazon S3 in the *Amazon VPC User Guide*.


Name : privateVPC
Add filter
1 to 1 of 1

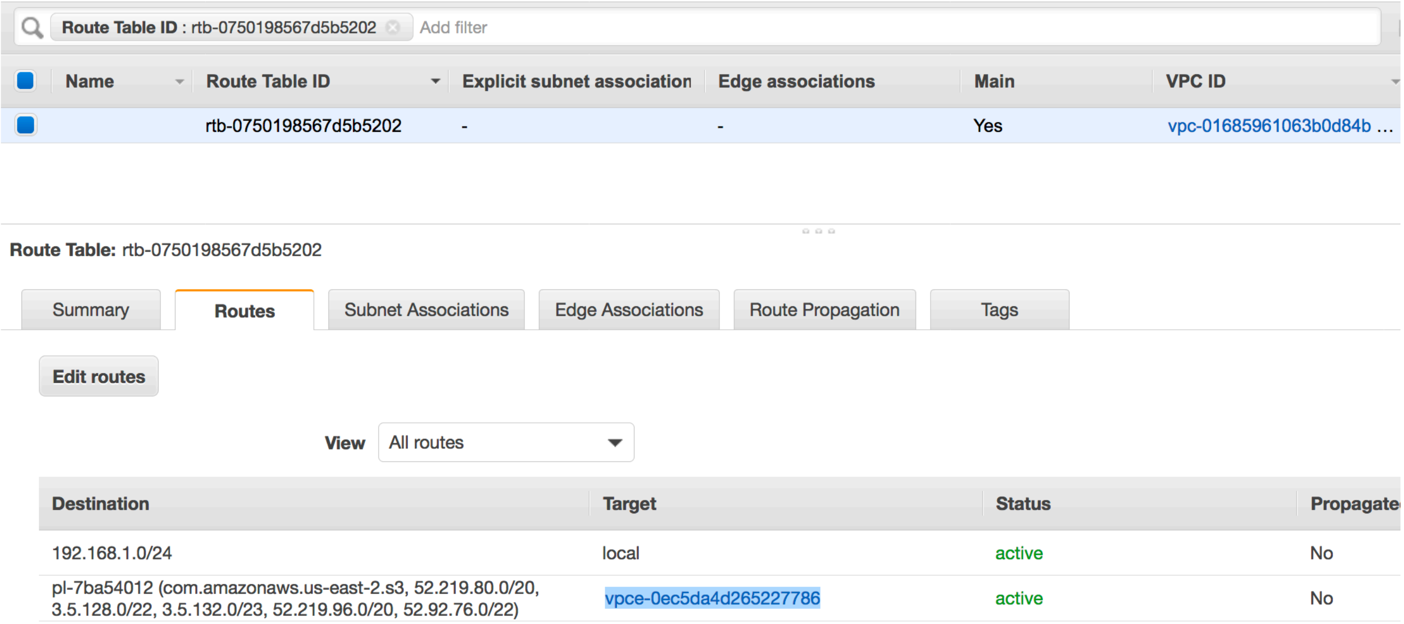
Name	VPC ID	State	IPv4 CIDR	IPv6	DHCP options se	Main Route table	Main Network ACL	Tenancy	Default VPC
privateVPC	vpc-01685961063b0d84b	available	192.168.1.0/24	-	dopt-a79e5acc	rtb-0750198567d5…	acl-02d197f2c9fbe46…	default	No

VPC: vpc-01685961063b0d84b

Description
CIDR Blocks
Flow Logs
Tags

VPC ID	vpc-01685961063b0d84b	**Tenancy**	default
State	available	**Default VPC**	No
IPv4 CIDR	192.168.1.0/24	**IPv6 CIDR**	-
IPv6 Pool	-	**DNS resolution**	Enabled
Network ACL	acl-02d197f2c9fbe46be	**DNS hostnames**	Disabled
DHCP options set	dopt-a79e5acc	**Route table**	rtb-0750198567d5b5202
Owner	261353713322		

- A route entry pointing to the VPC endpoint. For example vpce-0ec5da4d265227786 in the route table used by the VPC endpoint(vpce-0ec5da4d265227786).


Route Table ID : rtb-0750198567d5b5202
Add filter

Name	Route Table ID	Explicit subnet association	Edge associations	Main	VPC ID
	rtb-0750198567d5b5202	-	-	Yes	vpc-01685961063b0d84b …

Route Table: rtb-0750198567d5b5202

Summary
Routes
Subnet Associations
Edge Associations
Route Propagation
Tags
Edit routes

View All routes

Destination	Target	Status	Propagated
192.168.1.0/24	local	active	No
pl-7ba54012 (com.amazonaws.us-east-2.s3, 52.219.80.0/20, 3.5.128.0/22, 3.5.132.0/23, 52.219.96.0/20, 52.92.76.0/22)	vpce-0ec5da4d265227786	active	No

- A network ACL attached to the VPC allows the traffic.
- A security group attached to the VPC allows the traffic.

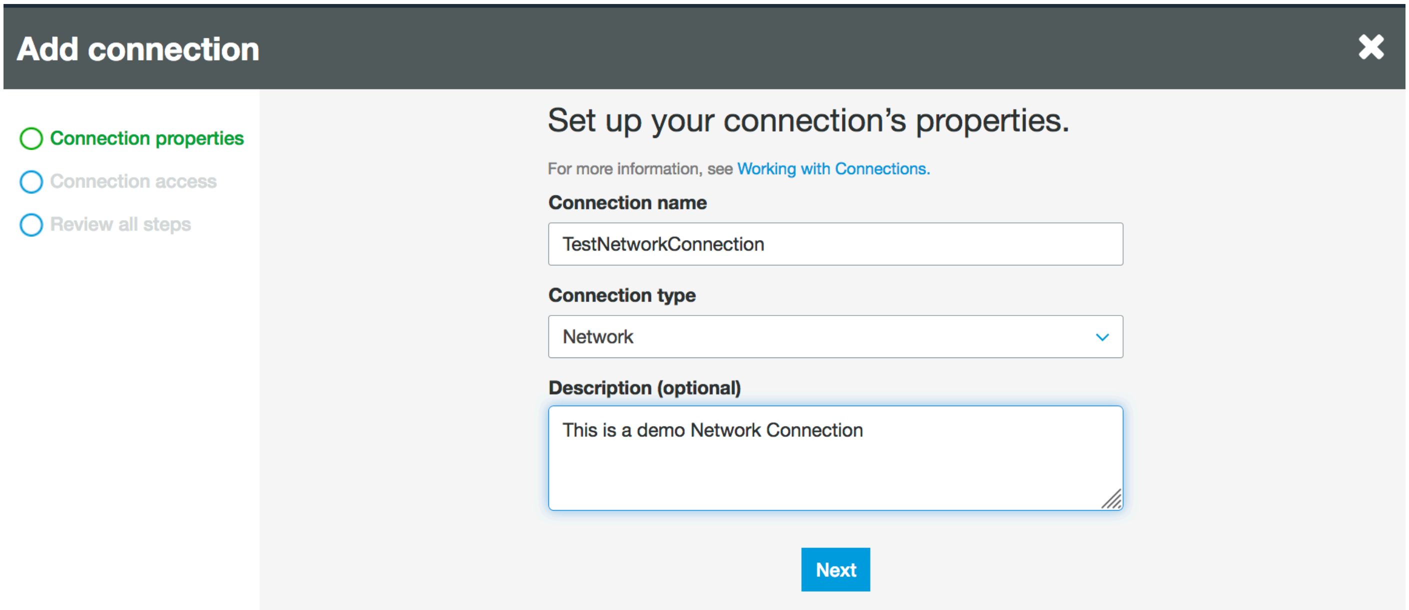
Creating the connection to Amazon S3

Typically, you create resources inside Amazon Virtual Private Cloud (Amazon VPC) so that they cannot be accessed over the public internet. By default, AWS Glue can't access resources inside a VPC. To enable AWS Glue to access resources inside your VPC, you must provide additional VPC-specific configuration information that includes VPC subnet IDs and security group IDs. To create a `Network` connection you need to specify the following information:

- A VPC ID
- A subnet within the VPC
- A security group

To set up a `Network` connection:

1. Choose **Add connection** in the navigation pane of the AWS Glue console.
2. Enter the connection name, choose **Network** as the connection type. Choose **Next**.



Add connection

Set up your connection's properties.

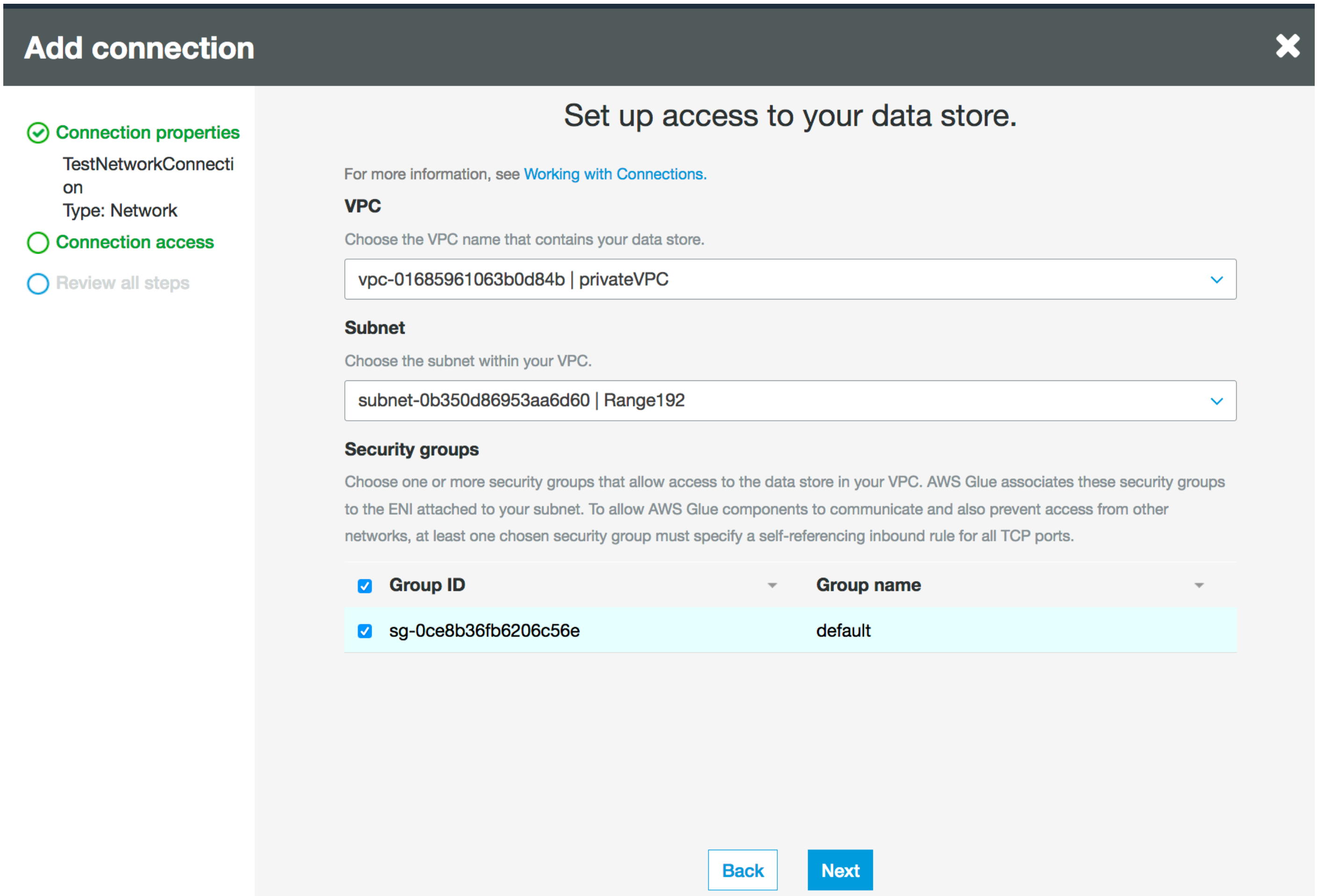
For more information, see Working with Connections.

Connection name: TestNetworkConnection

Connection type: Network

Description (optional): This is a demo Network Connection

3. Configure the VPC, Subnet and Security groups information.
   - VPC: choose the VPC name that contains your data store.
   - Subnet: choose the subnet within your VPC.
   - Security groups: choose one or more security groups that allow access to the data store in your VPC.



Add connection

Connection properties
TestNetworkConnection
Type: Network
Connection access
Review all steps

Set up access to your data store.

For more information, see Working with Connections.

VPC

Choose the VPC name that contains your data store.

vpc-01685961063b0d84b | privateVPC

Subnet

Choose the subnet within your VPC.

subnet-0b350d86953aa6d60 | Range192

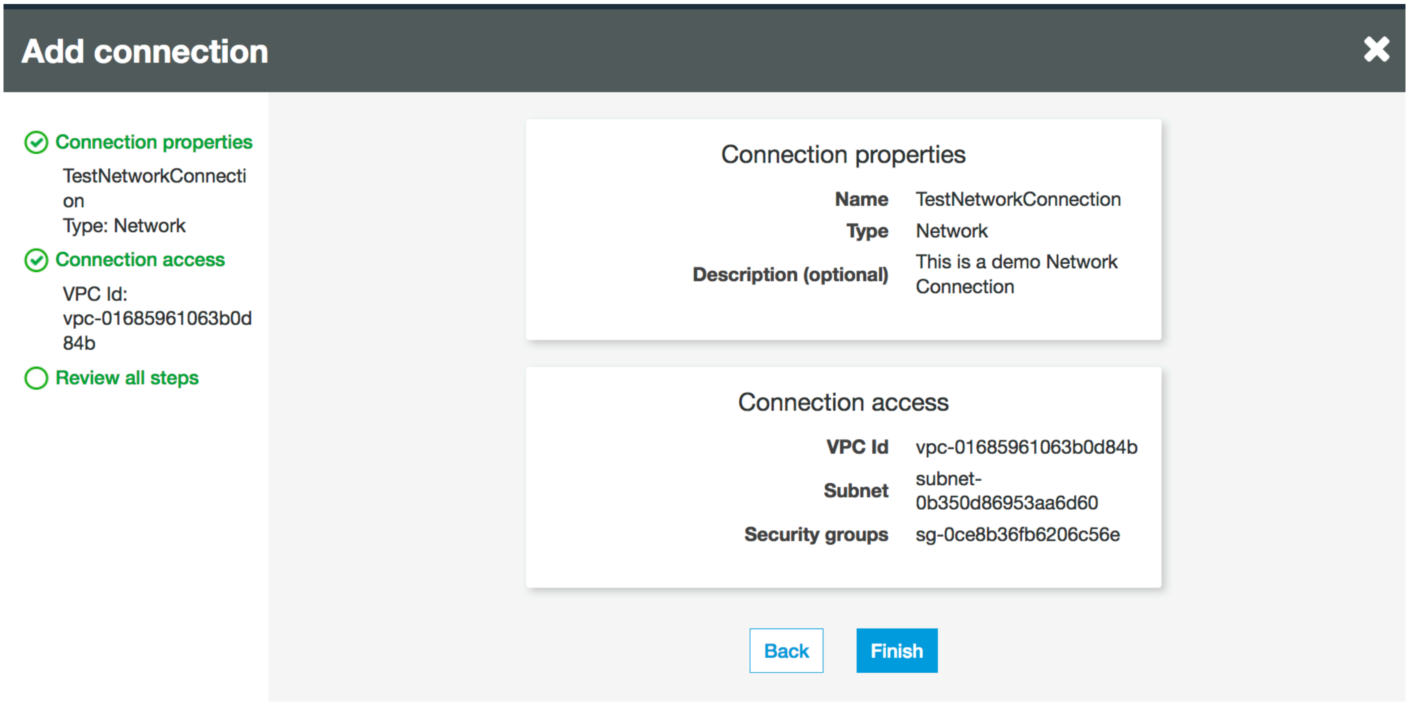
Security groups

Choose one or more security groups that allow access to the data store in your VPC. AWS Glue associates these security groups to the ENI attached to your subnet. To allow AWS Glue components to communicate and also prevent access from other networks, at least one chosen security group must specify a self-referencing inbound rule for all TCP ports.

☑ Group ID	Group name
☑ sg-0ce8b36fb6206c56e	default

4. Choose **Next**.

5. Verify the connection information and choose **Finish**.



Add connection

TestNetworkConnection
Type: Network

VPC Id: vpc-01685961063b0d84b

Connection properties

Name	TestNetworkConnection
Type	Network
Description (optional)	This is a demo Network Connection

Connection access

VPC Id	vpc-01685961063b0d84b
Subnet	subnet-0b350d86953aa6d60
Security groups	sg-0ce8b36fb6206c56e

Testing the connection to Amazon S3

Once you have created your `Network` connection, you can test the connectivity to your Amazon S3 data store in a VPC endpoint.

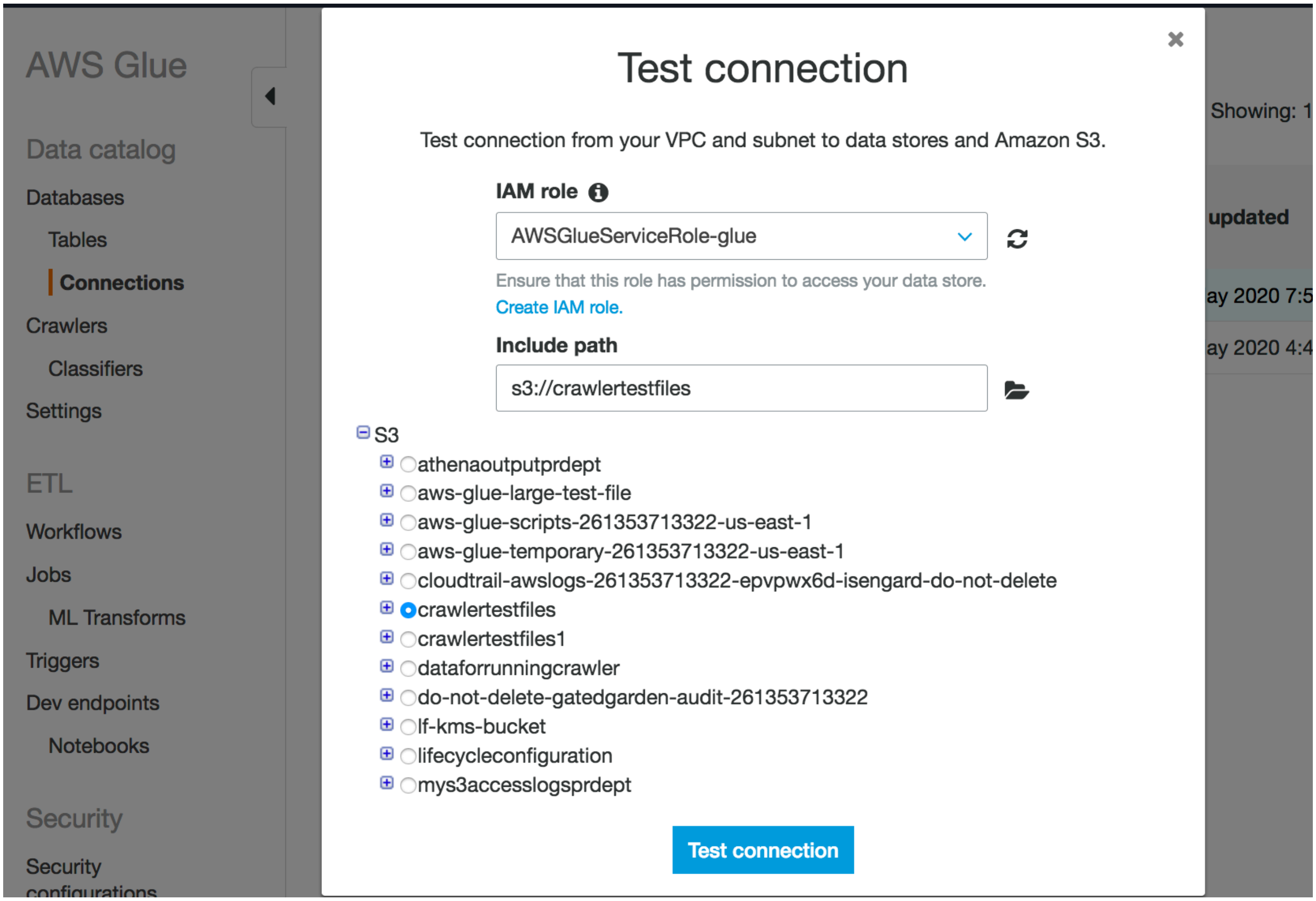
The following errors may occur when testing a connection:

- INTERNET CONNECTION ERROR: indicates an Internet connection issue
- INVALID BUCKET ERROR: indicates a problem with the Amazon S3 bucket
- S3 CONNECTION ERROR: indicates a failure to connect to Amazon S3
- INVALID CONNECTION TYPE: indicates the Connection type does not have the expected value, `NETWORK`
- INVALID CONNECTION TEST TYPE: indicates a problem with the type of network connection test
- INVALID TARGET: indicates that the Amazon S3 bucket has not been specified properly

To test a `Network` connection:

1. Select the **Network** connection in the AWS Glue console.
2. Choose **Test connection**.
3. Choose the IAM role that you created in the previous step and specify an Amazon S3 Bucket.

4. Choose **Test connection** to start the test. It might take few moments to show the result.


AWS Glue
Data catalog
Databases
Tables
Connections
Crawlers
Classifiers
Settings
ETL
Workflows
Jobs
ML Transforms
Triggers
Dev endpoints
Notebooks
Security
Security configurations
Showing: 1
updated
ay 2020 7:5
ay 2020 4:4
Test connection

Test connection

Test connection from your VPC and subnet to data stores and Amazon S3.

IAM role: AWSGlueServiceRole-glue

Ensure that this role has permission to access your data store.

Create IAM role.

Include path: s3://crawlertestfiles

S3
- ○ athenaoutputprdept
- ○ aws-glue-large-test-file
- ○ aws-glue-scripts-261353713322-us-east-1
- ○ aws-glue-temporary-261353713322-us-east-1
- ○ cloudtrail-awslogs-261353713322-epvpwx6d-isengard-do-not-delete
- ● crawlertestfiles
- ○ crawlertestfiles1
- ○ dataforrunningcrawler
- ○ do-not-delete-gatedgarden-audit-261353713322
- ○ lf-kms-bucket
- ○ lifecycleconfiguration
- ○ mys3accesslogsprdept

If you receive an error, check the following:

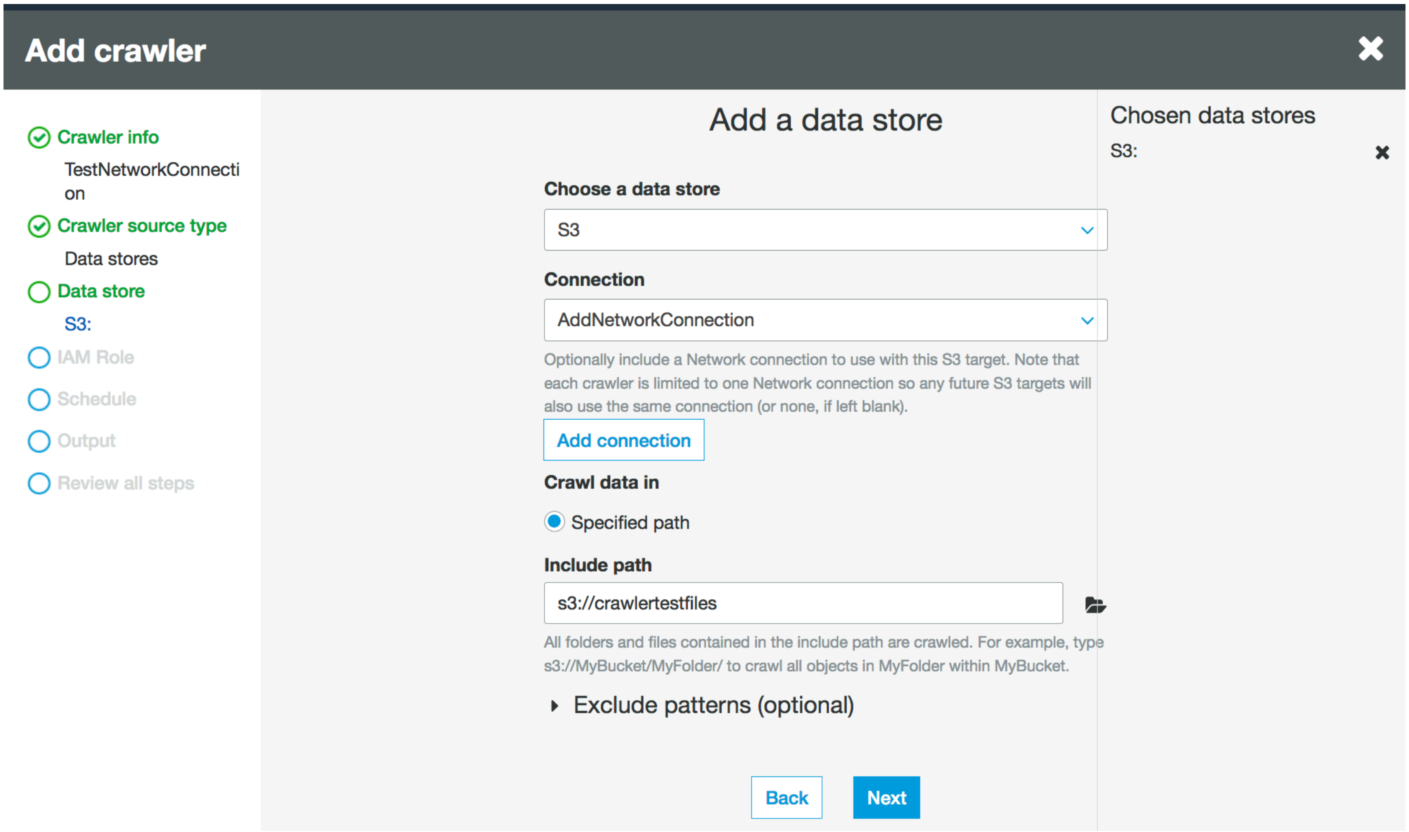
- The correct privileges are provided to the role selected.
- The correct Amazon S3 bucket is provided.
- The security groups and Network ACL allow the required incoming and outgoing traffic.
- The VPC you specified is connected to an Amazon S3 VPC endpoint.

Once you have successfully tested the connection, you can create a crawler.

Creating a crawler for an Amazon S3 data store

You can now create a crawler that specifies the `Network` connection you've created. For more details on creating a crawler, see Configuring a crawler.

1. Start by choosing **Crawlers** in the navigation pane on the AWS Glue console.
2. Choose **Add crawler**.
3. Specify the crawler name and choose **Next**.
4. When asked for the data source, choose **S3**, and specify the Amazon S3 bucket prefix and the connection you created earlier.



Add crawler

Crawler info: TestNetworkConnection

Crawler source type: Data stores

Data store: S3:

Add a data store

Choose a data store: S3

Connection: AddNetworkConnection

Optionally include a Network connection to use with this S3 target. Note that each crawler is limited to one Network connection so any future S3 targets will also use the same connection (or none, if left blank).

Crawl data in: ● Specified path

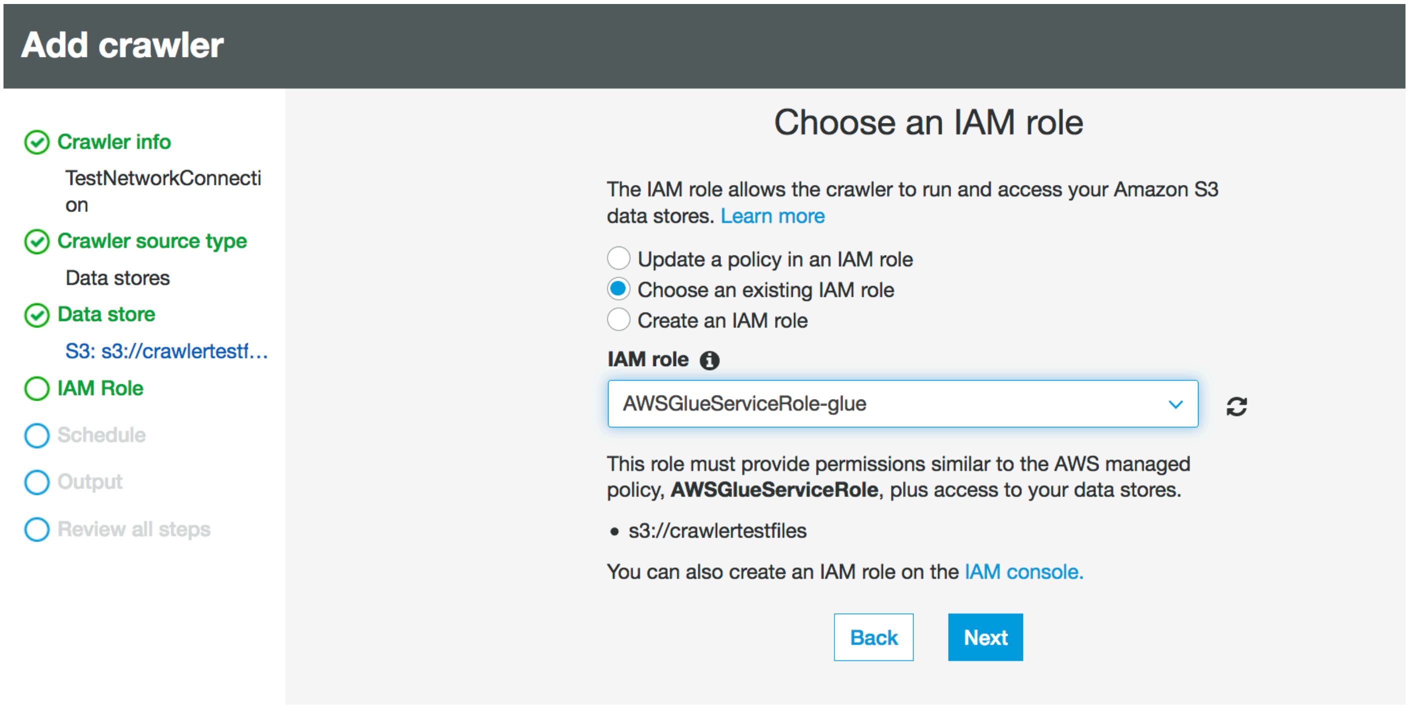
Include path: s3://crawlertestfiles

All folders and files contained in the include path are crawled. For example, type s3://MyBucket/MyFolder/ to crawl all objects in MyFolder within MyBucket.

Chosen data stores

S3:

5. If you need to, add another data store on the same network connection.
6. Choose IAM role. The IAM role must allow access to the AWS Glue service and the Amazon S3 bucket. For more information, see the section called “Configuring a crawler”.



Add crawler

Crawler info
TestNetworkConnection
Crawler source type
Data stores
Data store
S3: s3://crawlertestf…
IAM Role
Schedule
Output
Review all steps

Choose an IAM role

The IAM role allows the crawler to run and access your Amazon S3 data stores. Learn more

○ Update a policy in an IAM role
● Choose an existing IAM role
○ Create an IAM role

IAM role

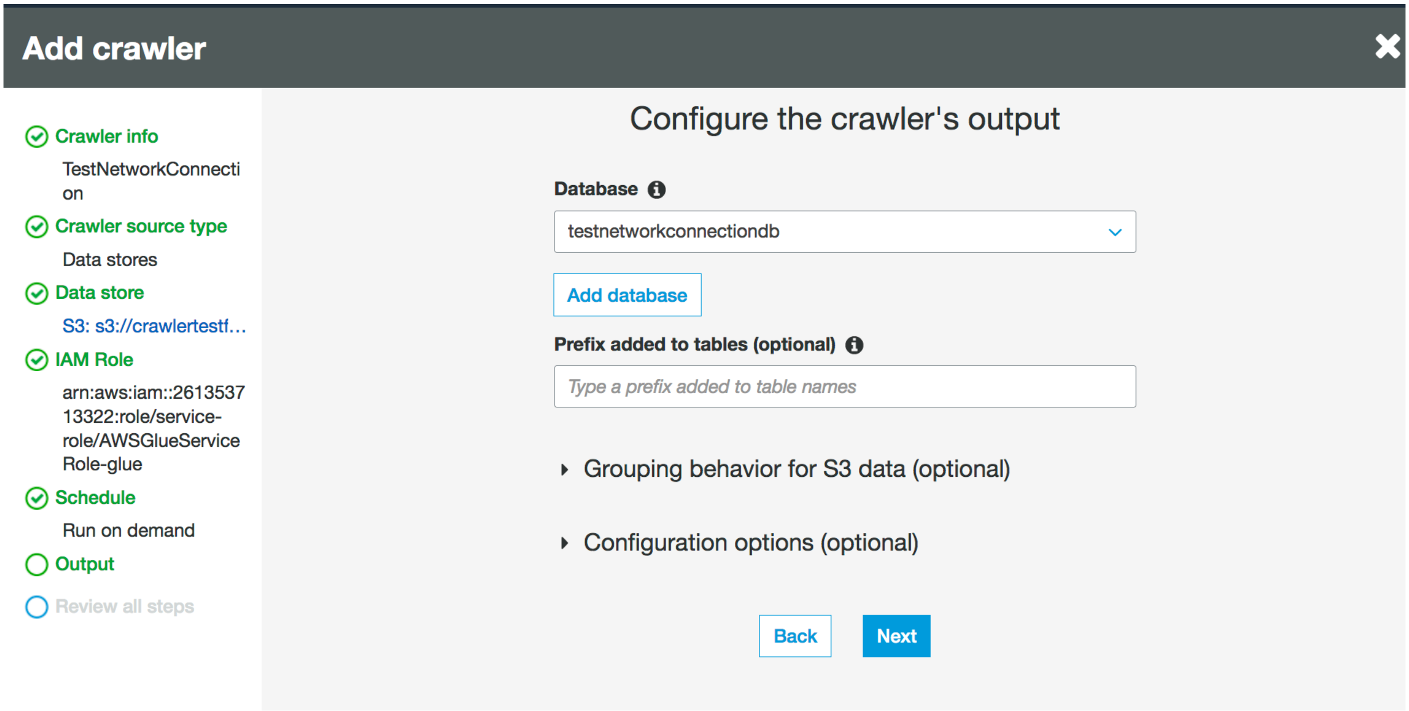
AWSGlueServiceRole-glue

This role must provide permissions similar to the AWS managed policy, **AWSGlueServiceRole**, plus access to your data stores.

- s3://crawlertestfiles

You can also create an IAM role on the IAM console.

7. Define the schedule for the crawler.
8. Choose an existing database in the Data Catalog, or create a new database entry.



Add crawler

Crawler info
TestNetworkConnection
Crawler source type
Data stores
Data store
S3: s3://crawlertestf…
IAM Role
arn:aws:iam::261353713322:role/service-role/AWSGlueServiceRole-glue
Schedule
Run on demand
Output
Review all steps

Configure the crawler's output

Database

testnetworkconnectiondb

Prefix added to tables (optional)

Type a prefix added to table names

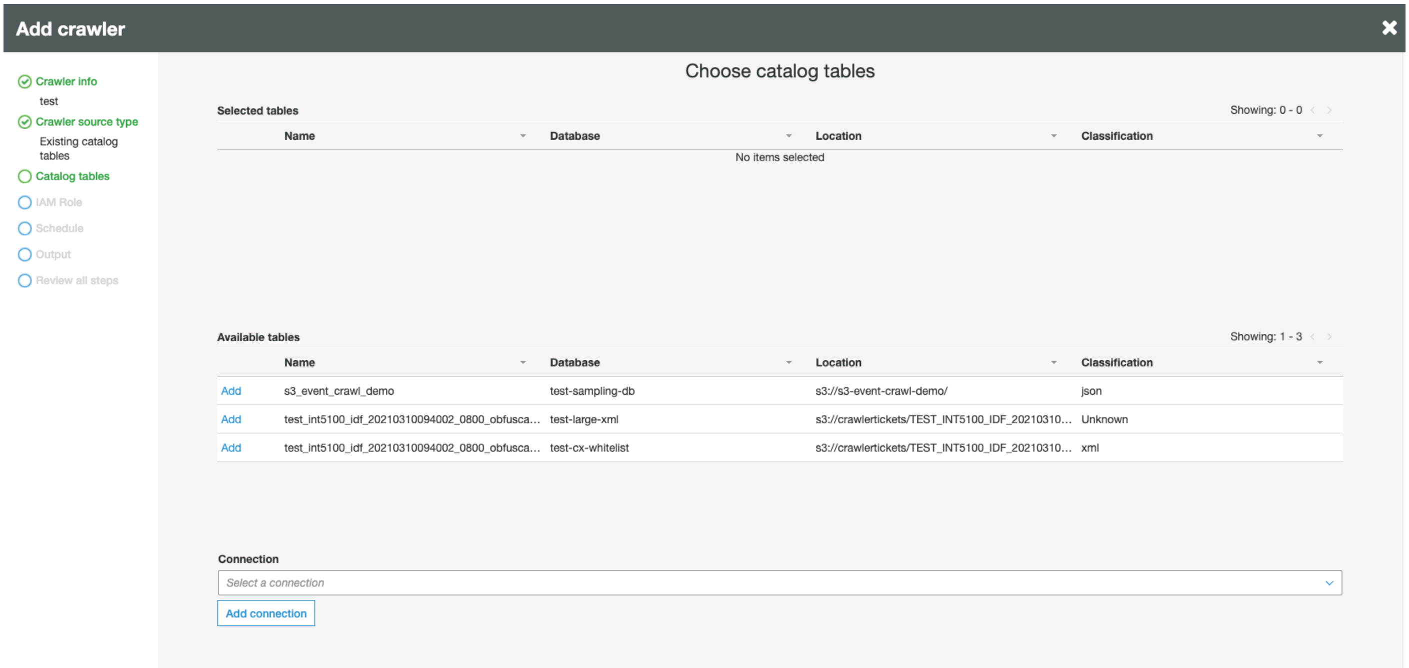
Grouping behavior for S3 data (optional)
Configuration options (optional)
Back
Next

9. Finish the remaining setup.

Creating a crawler for Amazon S3 backed Data Catalog tables

You can now create a crawler that specifies the `Network` connection you've created and a Catalog source type. For more details on creating a crawler, see Configuring a crawler.

1. Start by choosing **Crawlers** in the navigation pane on the AWS Glue console.
2. Choose **Add crawler**.
3. Specify the crawler name and choose **Next**.
4. When asked for the crawler source type, choose **Existing catalog tables**, and specify the existing catalog tables to crawl from the list of available tables.


Add crawler
Crawler info
test
Crawler source type
Existing catalog tables
Catalog tables
IAM Role
Schedule
Output
Review all steps
Choose catalog tables
Add connection

Selected tables Showing: 0 - 0

Name	Database	Location	Classification

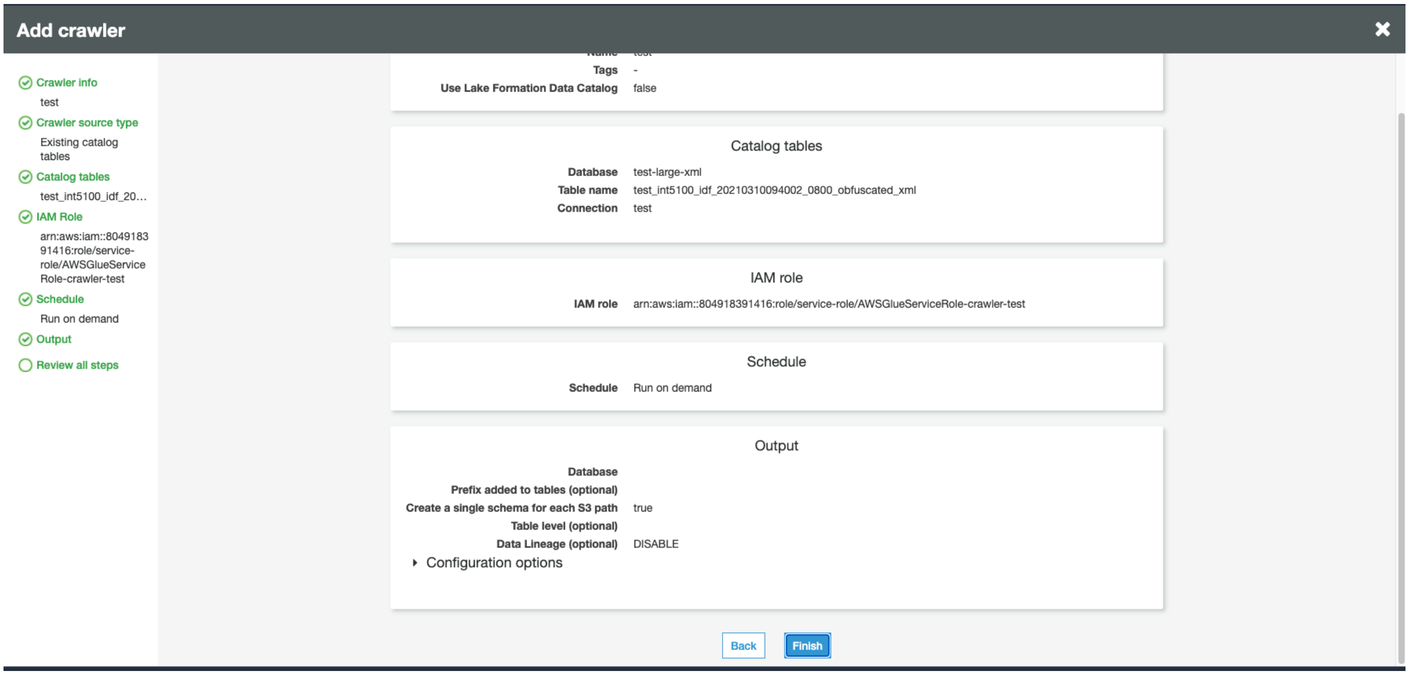
No items selected

Available tables Showing: 1 - 3

	Name	Database	Location	Classification
Add	s3_event_crawl_demo	test-sampling-db	s3://s3-event-crawl-demo/	json
Add	test_int5100_idf_20210310094002_0800_obfusca...	test-large-xml	s3://crawlertickets/TEST_INT5100_IDF_20210310...	Unknown
Add	test_int5100_idf_20210310094002_0800_obfusca...	test-cx-whitelist	s3://crawlertickets/TEST_INT5100_IDF_20210310...	xml

Connection: Select a connection

5. Choose IAM role. The IAM role must allow access to the AWS Glue service and the Amazon S3 bucket. For more information, see the section called “Configuring a crawler”.
6. Define the schedule for the crawler.
7. Choose an existing database in the Data Catalog, or create a new database entry.
8. Finish the remaining setup and review your steps.


Add crawler
Crawler info
test
Crawler source type
Existing catalog tables
Catalog tables
test_int5100_idf_20…
IAM Role
arn:aws:iam::804918391416:role/service-role/AWSGlueServiceRole-crawler-test
Schedule
Run on demand
Output
Review all steps

Tags -
Use Lake Formation Data Catalog false

Catalog tables

Database test-large-xml
Table name test_int5100_idf_20210310094002_0800_obfuscated_xml
Connection test

IAM role

IAM role arn:aws:iam::804918391416:role/service-role/AWSGlueServiceRole-crawler-test

Schedule

Schedule Run on demand

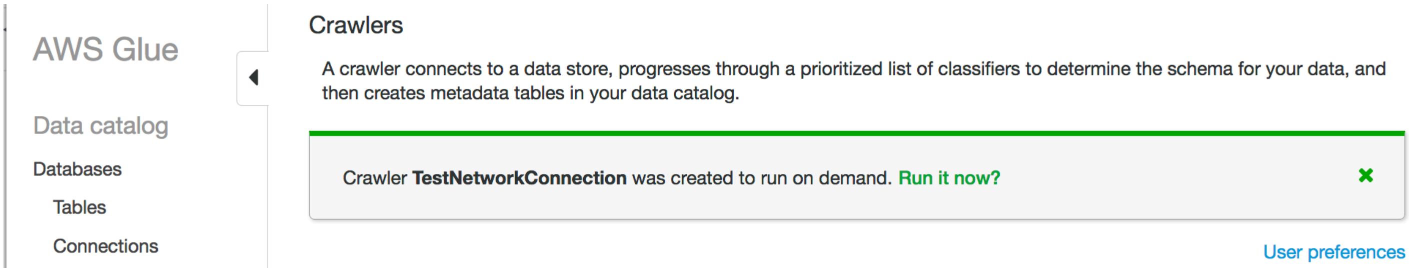
Output

Database
Prefix added to tables (optional)
Create a single schema for each S3 path true
Table level (optional)
Data Lineage (optional) DISABLE

Configuration options
Back
Finish

Running a crawler

Run your crawler.


AWS Glue
Data catalog
Databases
Tables
Connections
Crawlers
User preferences

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawler **TestNetworkConnection** was created to run on demand. Run it now?

Troubleshooting

For troubleshooting related to Amazon S3 buckets using a VPC gateway, see Why can't I connect to an S3 bucket using a gateway VPC endpoint?

Working with AWS Lake Formation-protected data

Full Table Access

AWS Glue supports Full Table Access (FTA) for AWS Lake Formation-protected tables. This allows your ETL jobs to read and write data with full table permissions.

Prerequisites

- Appropriate IAM roles and permissions
- AWS Lake Formation configured for your data catalog
- Compatible table types (Hive or Iceberg)

Key considerations

Required permissions

- `lakeformation:GetDataAccess` IAM permission
- AWS Lake Formation table permissions
- Amazon S3 bucket access permissions

Supported table types

- Hive tables
- Iceberg tables

Limitations

- Not compatible with Spark Streaming
- Cannot be used simultaneously with fine-grained access control
- Does not support Delta or Hudi tables

Best Practices

1. Ensure IAM roles are properly configured before job execution
2. Test access permissions in a development environment
3. Monitor job execution logs for permission-related issues
4. Maintain clear documentation of access patterns

Troubleshooting

Common issues include:

- Missing IAM permissions
- Incorrect AWS Lake Formation configuration
- Table type compatibility problems

For complete setup instructions and configuration details, see Using AWS Lake Formation with full table access.

Troubleshooting connection issues in AWS Glue

When an AWS Glue crawler or a job uses connection properties to access a data store, you might encounter errors when you try to connect. AWS Glue uses private IP addresses in the subnet when it creates elastic network interfaces in your specified virtual private cloud (VPC) and subnet. Security groups specified in the connection are applied on each of the elastic network interfaces. Check to see whether security groups allow outbound access and if they allow connectivity to the database cluster.

In addition, Apache Spark requires bidirectional connectivity among driver and executor nodes. One of the security groups needs to allow ingress rules on all TCP ports. You can prevent it from being open to the world by restricting the source of the security group to itself with a self-referencing security group.

Here are some typical actions you can take to troubleshoot connection problems:

- Check the port address of your connection.
- Check the user name and password string in your connection or secret.
- For a JDBC data store, verify that it allows incoming connections.
- Verify that your data store can be accessed within your VPC.
- If you store your connection credentials using AWS Secrets Manager, make sure that your IAM role for AWS Glue has permission to access your secret. For more information, see Example: Permission to retrieve secret values in the *AWS Secrets Manager User Guide*. Depending on your network setup, you might also need to create a VPC endpoint to establish a private connection between your VPC and Secrets Manager. For more information, see Using an AWS Secrets Manager VPC endpoint.

Tutorial: Using the AWS Glue Connector for Elasticsearch

Elasticsearch is a popular open-source search and analytics engine for use cases such as log analytics, real-time application monitoring, and clickstream analysis. You can use OpenSearch as a data store for your extract, transform, and load (ETL) jobs by configuring the AWS Glue Connector for Elasticsearch in AWS Glue Studio. This connector is available for free from AWS Marketplace.

Note

The AWS Marketplace Elasticsearch Spark Connector has been deprecated. Please use the AWS Glue Connector for Elasticsearch instead.

In this tutorial, we will show how to connect to your Amazon OpenSearch Service nodes with a minimal number of steps.

Topics

- Prerequisites
- Step 1: (Optional) Create an AWS secret for your OpenSearch cluster information
- Step 2: Subscribe to the connector
- Step 3: Activate the connector in AWS Glue Studio and create a connection
- Step 4: Configure an IAM role for your ETL job
- Step 5: Create a job that uses the OpenSearch connection
- Step 6: Run the job

Prerequisites

To use this tutorial, you must have the following:

- Access to AWS Glue Studio
- Access to an OpenSearch cluster in the AWS Cloud
- (Optional) Access to AWS Secrets Manager.

Step 1: (Optional) Create an AWS secret for your OpenSearch cluster information

To safely store and use your connection credential, save your credential in AWS Secrets Manager. The secret you create will be used later in the tutorial by the connection. The credential key-value pairs will be fed into the AWS Glue Connector for Elasticsearch as normal connection options.

For more information about creating secrets, see Creating and Managing Secrets with AWS Secrets Manager in the *AWS Secrets Manager User Guide*.

To create an AWS secret

1. Sign in to the AWS Secrets Manager console.
2. On either the service introduction page or the **Secrets** list page, choose **Store a new secret**.
3. On the **Store a new secret** page, choose **Other type of secret**. This option means that you must supply the structure and details of your secret.
4. Add a **Key** and **Value** pair for the OpenSearch cluster user name. For example:

   `es.net.http.auth.user`: *`username`*
5. Choose **+ Add row**, and enter another key-value pair for the password. For example:

   `es.net.http.auth.pass`: *`password`*
6. Choose **Next**.
7. Enter a secret name. For example: **my-es-secret**. You can optionally include a description.

   Record the secret name, which is used later in this tutorial, and then choose **Next**.
8. Choose **Next** again, and then choose **Store** to create the secret.

Next step

Step 2: Subscribe to the connector

Step 2: Subscribe to the connector

The AWS Glue Connector for Elasticsearch is available for free from AWS Marketplace.

To subscribe to the AWS Glue Connector for Elasticsearch on AWS Marketplace

1. If you have not already configured your AWS account to use License Manager, do the following:
    a. Open the AWS License Manager console at https://console.aws.amazon.com/license-manager.
    b. Choose **Create customer managed license**.
    c. In the **IAM permissions (one-time setup)** window, choose **I grant AWS License Manager the required permissions**, and then choose **Grant permissions**.

       If you do not see this window, then you have already configured the necessary permissions.
2. Open the AWS Glue Studio console at https://console.aws.amazon.com/gluestudio/.
3. In the AWS Glue Studio console, expand the menu icon (☰), and then choose **Connectors** in the navigation pane.
4. On the **Connectors** page, choose **Go to AWS Marketplace**.
5. In AWS Marketplace, in the **Search AWS Glue Studio products** section, enter **AWS Glue Connector for Elasticsearch** in the search field, and then press Enter.
6. Choose the name of the connector, **AWS Glue Connector for Elasticsearch**.
7. On the product page for the connector, use the tabs to view information about the connector. When you're ready to continue, choose **Continue to Subscribe**.
8. Review the terms of use for the software. Click **Accept Terms**.
9. When the subscription process completes, you will see a notification: "Thank you for subscribing to this product! You can now configure your software." Above the banner will be the button **Continue to Configuration**. Choose **Continue to Configuration**.
10. Choose the Fulfillment option on the **Configure this software** page. You can either choose between AWS Glue 1.0/2.0 or AWS Glue 3.0. Then, choose **Continue to Launch**.

Next step

[Step 3: Activate the connector in AWS Glue Studio and create a connection](#)

Step 3: Activate the connector in AWS Glue Studio and create a connection

After you choose **Continue to Launch**, you see the **Launch this software** page in AWS Marketplace. After you use the link to activate the connector in AWS Glue Studio, you create a connection.

To deploy the connector and create a connection in AWS Glue Studio

1. On the **Launch this software** page in the AWS Marketplace console, choose **Usage Instructions**, and then choose the link in the window that appears.

   Your browser is redirected to the AWS Glue Studio console **Create marketplace connection** page.
2. Enter a name for the connection. For example: **my-es-connection**.
3. In the **Connection access** section, for **Connection credential type**, choose **User name and password**.
4. For the **AWS secret**, enter the name of your secret. For example: **my-es-secret**.
5. In the **Network options** section, enter the VPC information to connect to OpenSearch cluster.
6. Choose **Create connection and activate connector**.

Next step

Step 4: Configure an IAM role for your ETL job

Step 4: Configure an IAM role for your ETL job

When you create the AWS Glue ETL job, you specify an AWS Identity and Access Management (IAM) role for the job to use. The role must grant access to all resources used by the job, including Amazon S3 (for any sources, targets, scripts, driver files, and temporary directories), and also AWS Glue Data Catalog objects.

The assumed IAM role for the AWS Glue ETL job must also have access to the secret that was created in the previous section. By default, the AWS managed role `AWSGlueServiceRole` does not have access to the secret. To set up access control for your secrets, see Authentication and Access Control for AWS Secrets Manager and Limiting Access to Specific Secrets.

To configure an IAM role for your ETL job

1. Configure the permissions described in the section called “Review IAM permissions needed for ETL jobs”.
2. Configure the additional permissions needed when using connectors with AWS Glue Studio, as described in the section called “Permissions required for using connectors”.

Next step

Step 5: Create a job that uses the OpenSearch connection

Step 5: Create a job that uses the OpenSearch connection

After creating a role for your ETL job, you can create a job in AWS Glue Studio that uses the connection and connector for Open Spark ElasticSearch.

If your job runs within a Amazon Virtual Private Cloud (Amazon VPC), make sure the VPC is configured correctly. For more information, see the section called “Configure a VPC for your ETL job”.

To create a job that uses the Elasticsearch Spark Connector

1. In AWS Glue Studio, choose **Connectors**.
2. In the **Your connections** list, select the connection you just created and choose **Create job**.
3. In the visual job editor, choose the Data source node. On the right, on the **Data source properties - Connector** tab, configure additional information for the connector.

    a. Choose **Add schema** and enter the schema of the data set in the data source. Connections do not use tables stored in the Data Catalog, which means that AWS Glue Studio doesn't know the schema of the data. You must manually provide this schema information. For instructions on how to use the schema editor, see the section called “Editing the schema in a custom transform node”.

    b. Expand **Connection options**.

    c. Choose **Add new option** and enter the information needed for the connector that was not entered in the AWS secret:

    - **es.nodes**: https://*<OpenSearch domain endpoint>*
    - **es.port**: 443

  - **path**: test
  - **es.nodes.wan.only**: true

  For an explanation of these connection options, refer to: https://www.elastic.co/guide/en/elasticsearch/hadoop/current/configuration.html.

4. Add a target node to the graph.

   Your data target can be Amazon S3, or it can use information from an AWS Glue Data Catalog or a connector to write data in a different location. For example, you can use a Data Catalog table to write to a database in Amazon RDS, or you can use a connector as your data target to write to data stores that are not natively supported in AWS Glue.

   If you choose a connector for your data target, you must choose a connection created for that connector. Also, if required by the connector provider, you must add options to provide additional information to the connector. If you use a connection that contains information for an AWS secret, then you don't need to provide the user name and password authentication in the connection options.
5. Optionally, add additional data sources and one or more transform nodes as described in the section called "Transform data with AWS Glue managed transforms".
6. Configure the job properties as described in the section called "Modify the job properties", starting with step 3, and save the job.

Next step

Step 6: Run the job

Step 6: Run the job

After you save your job, you can run the job to perform the ETL operations.

To run the job you created for the AWS Glue Connector for Elasticsearch

1. Using the AWS Glue Studio console, on the visual editor page, choose **Run**.
2. In the success banner, choose **Run Details**, or you can choose the **Runs** tab of the visual editor to view information about the job run.

Building AWS Glue jobs with interactive sessions

Data engineers can author AWS Glue jobs faster and more easily than before using interactive sessions in AWS Glue.

Topics

- Overview of AWS Glue interactive sessions
- Getting started with AWS Glue interactive sessions
- Configuring AWS Glue interactive sessions for Jupyter and AWS Glue Studio notebooks
- Converting a script or notebook into an AWS Glue job
- Working with streaming operations in AWS Glue interactive sessions
- Developing and testing AWS Glue job scripts locally
- Development endpoints

Overview of AWS Glue interactive sessions

With AWS Glue interactive sessions, you can rapidly build, test, and run data preparation and analytics applications. Interactive sessions provides a programmatic and visual interface for building and testing extract, transform, and load (ETL) scripts for data preparation. Interactive sessions run Apache Spark analytics applications and provide on-demand access to a remote Spark runtime environment. AWS Glue transparently manages serverless Spark for these interactive sessions.

Interactive sessions are flexible, so you build and test your applications from the environment of your choice. You can create and work with interactive sessions through the AWS Command Line Interface and the API. You can use Jupyter-compatible notebooks to visually author and test your notebook scripts. Interactive sessions provide an open-source Jupyter kernel that integrates almost anywhere that Jupyter does, including integrating with IDEs such as PyCharm, IntelliJ, and VS Code. This enables you to author code in your local environment and run it seamlessly on the interactive sessions backend.

Using the interactive sessions API, customers can programmatically run applications that use Apache Spark analytics without having to manage Spark infrastructure. You can run one or more Spark statements within a single interactive session.

Interactive sessions therefore provide a faster, cheaper, more-flexible way to build and run data preparation and analytics applications. To learn how to use interactive sessions, see the documentation in this section. Magics supported by AWS Glue

Limitations

- Job bookmarks are not supported in interactive sessions.
- Creating notebook jobs using the AWS Command Line Interface is not supported.
- AWS Glue Studio notebooks do not support Scala.

Getting started with AWS Glue interactive sessions

These sections describe how to run AWS Glue interactive sessions locally.

Prerequisites for setting up interactive sessions locally

The following are prerequisites for installing interactive sessions:

- Supported Python versions are 3.6 - 3.10+.
- See sections below for MacOS/Linux and Windows instructions.

Installing Jupyter and AWS Glue interactive sessions Jupyter kernels

Use the following to install the kernel locally.

The command, `install-glue-kernels`, installs the jupyter kernelspec for both pyspark and spark kernels and also installs logos in the right directory.

```
pip3 install --upgrade jupyter boto3 aws-glue-sessions
```

```
install-glue-kernels
```

Running Jupyter

To run Jupyter Notebook, complete the following steps.

1. Run the following command to launch Jupyter Notebook.

   ```
   jupyter notebook
   ```

2. Choose **New**, and then choose one of the AWS Glue kernels to begin coding against AWS Glue.

Configuring session credentials and region

MacOS/Linux instructions

AWS Glue interactive sessions requires the same IAM permissions as AWS Glue Jobs and Dev Endpoints. Specify the role used with interactive sessions in one of two ways:

1. With the `%iam_role` and `%region` magics
2. With an additional line in `~/.aws/config`

Configuring a session role with magic

In the first cell, type `%iam_role <YourGlueServiceRole>` in the first cell executed.

Configuring a session role with `~/.aws/config`

AWS Glue Service Role for interactive sessions can either be specified in the notebook itself or stored alongside the AWS CLI config. If you have a role you typically use with AWS Glue Jobs this will be that role. If you do not have a role you use for AWS Glue jobs, please follow this guide, Configuring IAM permissions for AWS Glue , to set one up.

To set this role as the default role for interactive sessions:

1. With a text editor, open `~/.aws/config`.
2. Look for the profile you use for AWS Glue. If you don't use a profile, use the `[Default]` profile.
3. Add a line in the profile for the role you intend to use like `glue_role_arn=<AWSGlueServiceRole>`.
4. [Optional]: If your profile does not have a default region set, I recommend adding one with `region=us-east-1`, replacing `us-east-1` with your desired region.
5. Save the config.

For more information, see Interactive sessions with IAM.

Windows instructions

AWS Glue interactive sessions requires the same IAM permissions as AWS Glue Jobs and Dev Endpoints. Specify the role used with interactive sessions in one of two ways:

1. With the `%iam_role` and `%region` magics
2. With an additional line in `~/.aws/config`

Configuring a session role with magic

In the first cell, type `%iam_role <YourGlueServiceRole>` in the first cell executed.

Configuring a session role with `~/.aws/config`

AWS Glue Service Role for interactive sessions can either be specified in the notebook itself or stored alongside the AWS CLI config. If you have a role you typically use with AWS Glue Jobs this will be that role. If you do not have a role you use for AWS Glue jobs, please follow this guide, Setting up IAM permissions for AWS Glue , to set one up.

To set this role as the default role for interactive sessions:

1. With a text editor, open `~/.aws/config`.
2. Look for the profile you use for AWS Glue. If you don't use a profile, use the `[Default]` profile.
3. Add a line in the profile for the role you intend to use like `glue_role_arn=<AWSGlueServiceRole>`.
4. [Optional]: If your profile does not have a default region set, I recommend adding one with `region=us-east-1`, replacing `us-east-1` with your desired region.
5. Save the config.

For more information, see Interactive sessions with IAM.

Upgrading from the interactive sessions preview

The kernel was upgraded with new names when it was released with version 0.27. To clean up preview versions of the kernels run the following from a terminal or PowerShell.

Note

If you are a part of any other AWS Glue preview that requires a custom service model, removing the kernel will remove the custom service model.

```
# Remove Old Glue Kernels
jupyter kernelspec remove glue_python_kernel
jupyter kernelspec remove glue_scala_kernel

# Remove Custom Model
cd ~/.aws/models
rm -rf glue/
```

Using interactive sessions with SageMaker AI Studio

AWS Glue Interactive Sessions is an on-demand, serverless, Apache Spark runtime environment that data scientists and engineers can use to rapidly build, test, and run data preparation and analytics applications. You can initiate an AWS Glue interactive session by starting a Amazon SageMaker AI Studio Classic notebook.

For more information, see Prepare Data using AWS Glue interactive sessions .

Using interactive sessions with Microsoft Visual Studio Code

Prerequisites

- Install AWS Glue interactive sessions and verify it works with Jupyter Notebook.
- Download and install Visual Studio Code with Jupyter. For details, see Jupyter Notebook in VS Code.

To get started with interactive sessions with VSCode

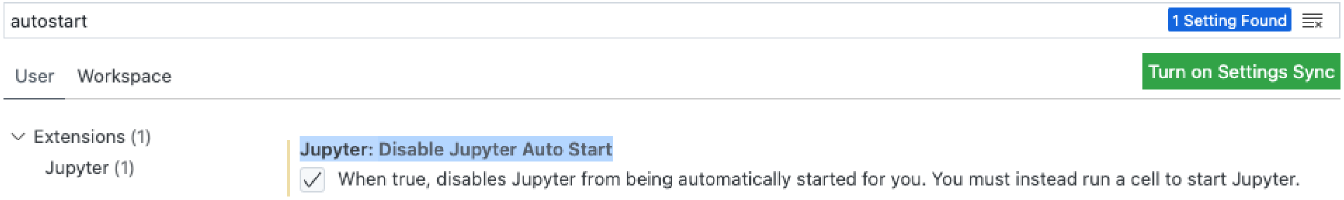
1. Disable Jupyter AutoStart in VS Code.

   In Visual Studio Code, Jupyter kernels will auto-start which will prevent your magics from taking effect as the session will already be started. To disable **Auto Start** on Windows, go to

File > **Preferences** > **Extensions** > **Jupyter** > right-click on Jupyter then choose **Extension Settings**.

On MacOS, go to **Code** > **Settings** > **Extensions** > **Jupyter** > right-click on Jupyter then choose **Extension Settings**.

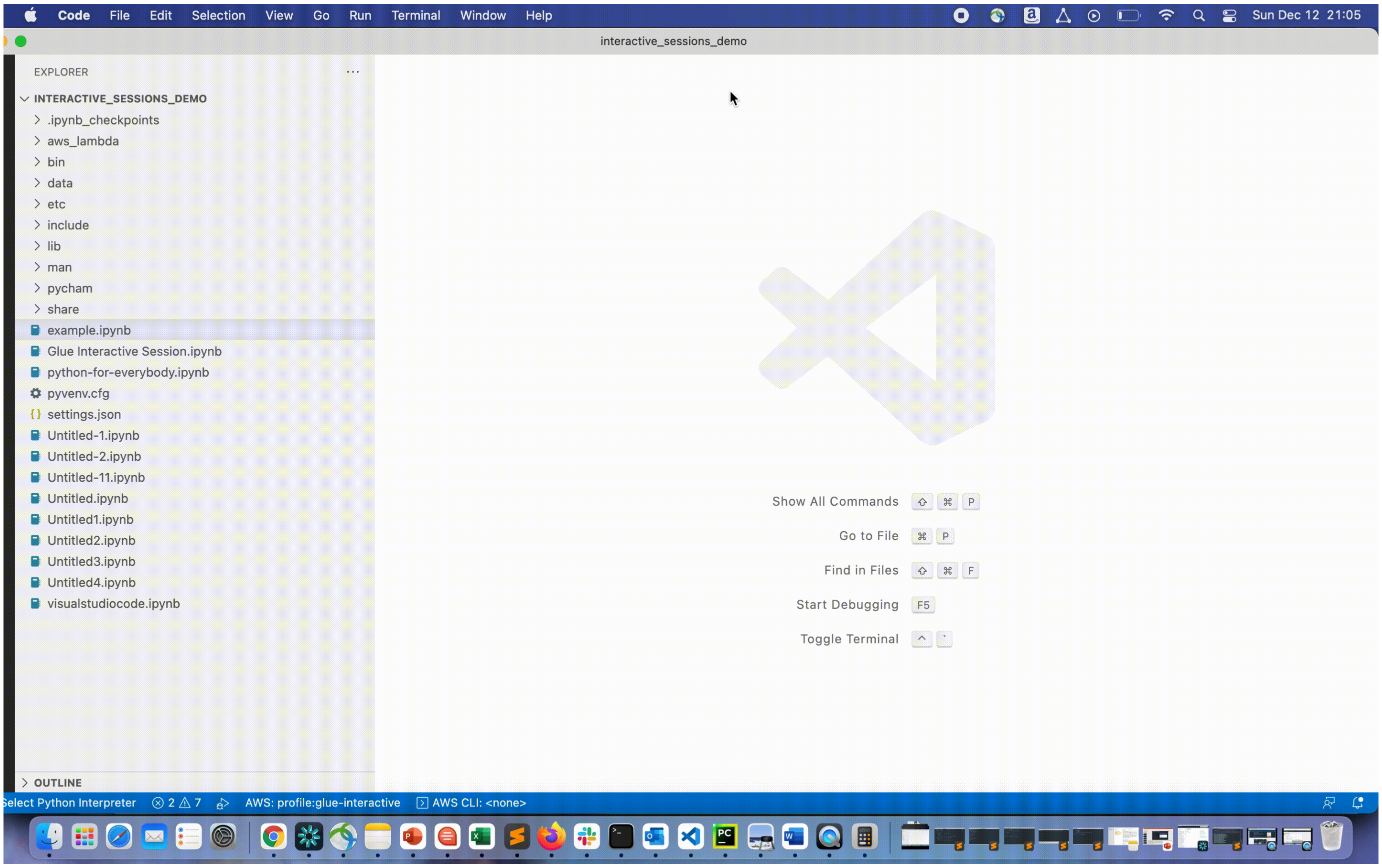
Scroll down until you see **Jupyter: Disable Jupyter Auto Start**. Check the box "When true, disables Jupyter from being automatically started for you. You must instead run a cell to start Jupyter."


autostart
1 Setting Found
User
Workspace
Turn on Settings Sync
Extensions (1)
Jupyter (1)

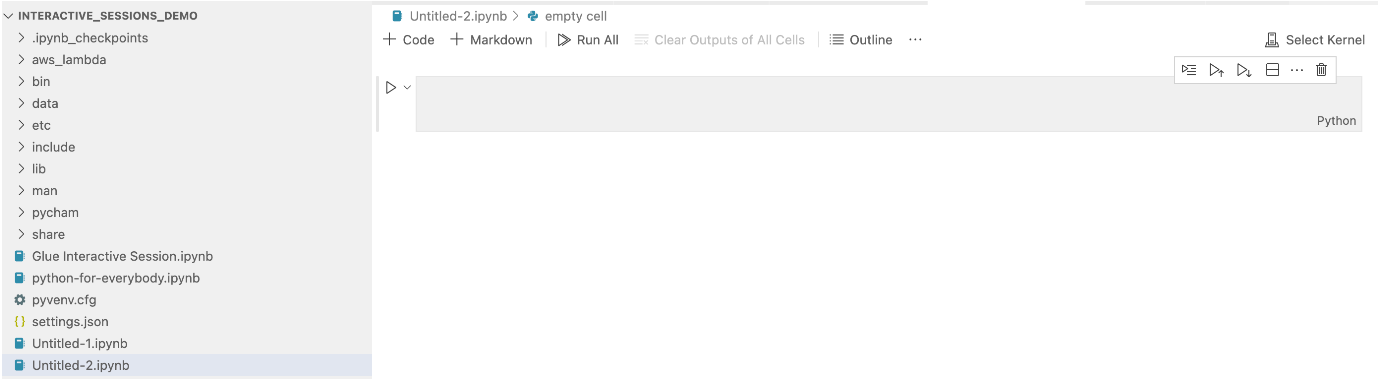
Jupyter: Disable Jupyter Auto Start

☑ When true, disables Jupyter from being automatically started for you. You must instead run a cell to start Jupyter.

2. Go to File > New File > Save to save this file with name of your choice as an `.ipynb` extension or select **jupyter** under **select a language** and save the file.


Code
File
Edit
Selection
View
Go
Run
Terminal
Window
Help
Sun Dec 12 21:05
interactive_sessions_demo
EXPLORER
INTERACTIVE_SESSIONS_DEMO
.ipynb_checkpoints
aws_lambda
bin
data
etc
include
lib
man
pycham
share
example.ipynb
Glue Interactive Session.ipynb
python-for-everybody.ipynb
pyvenv.cfg
settings.json
Untitled-1.ipynb
Untitled-2.ipynb
Untitled-11.ipynb
Untitled.ipynb
Untitled1.ipynb
Untitled2.ipynb
Untitled3.ipynb
Untitled4.ipynb
visualstudiocode.ipynb
Show All Commands ⇧ ⌘ P
Go to File ⌘ P
Find in Files ⇧ ⌘ F
Start Debugging F5
Toggle Terminal ^ `
OUTLINE
Select Python Interpreter
⊗ 2 ⚠ 7
AWS: profile:glue-interactive
AWS CLI: <none>

3. Double-click on the file. The Jupyter shell will display and a notebook will be opened.


INTERACTIVE_SESSIONS_DEMO
.ipynb_checkpoints
aws_lambda
bin
data
etc
include
lib
man
pycham
share
Glue Interactive Session.ipynb
python-for-everybody.ipynb
pyvenv.cfg
settings.json
Untitled-1.ipynb
Untitled-2.ipynb
Untitled-2.ipynb > empty cell
+ Code
+ Markdown
Run All
Clear Outputs of All Cells
Outline
Select Kernel
Python

4. On Windows, when you first create a file, by default no kernel is selected. Click on **Select Kernel** and a list of available kernels is displayed. Choose **Glue PySpark**.

   On MacOS, If you do not see the **Glue PySpark** kernel, try the following steps:

   1. Run a local Jupyter session to obtain the URL.

      For example, run the following command to launch Jupyter Notebook.

      ```
      jupyter notebook
      ```

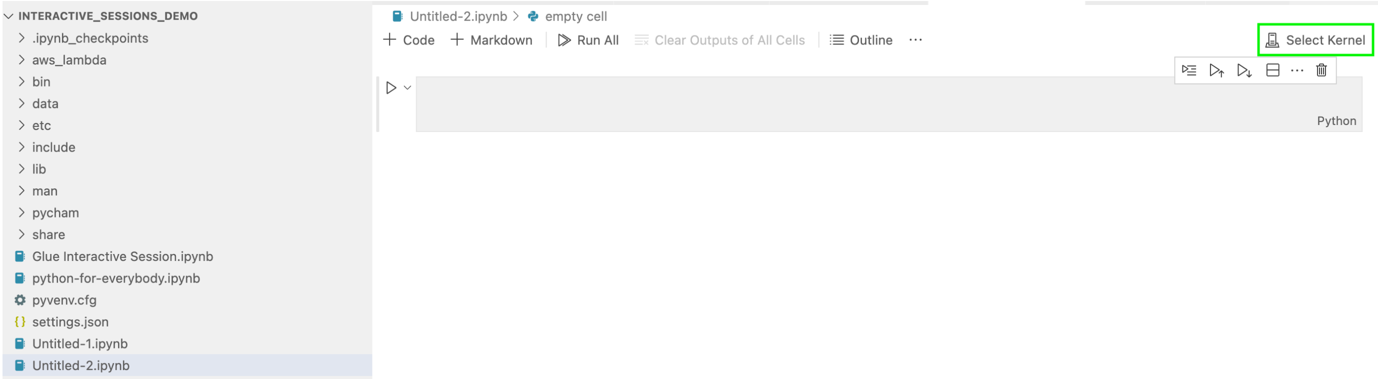
      When the notebook first runs, you will see a URL that looks like `http://localhost:8888/?token=3398XXXXXXXXXXXXXXXX`.

      Copy the URL.

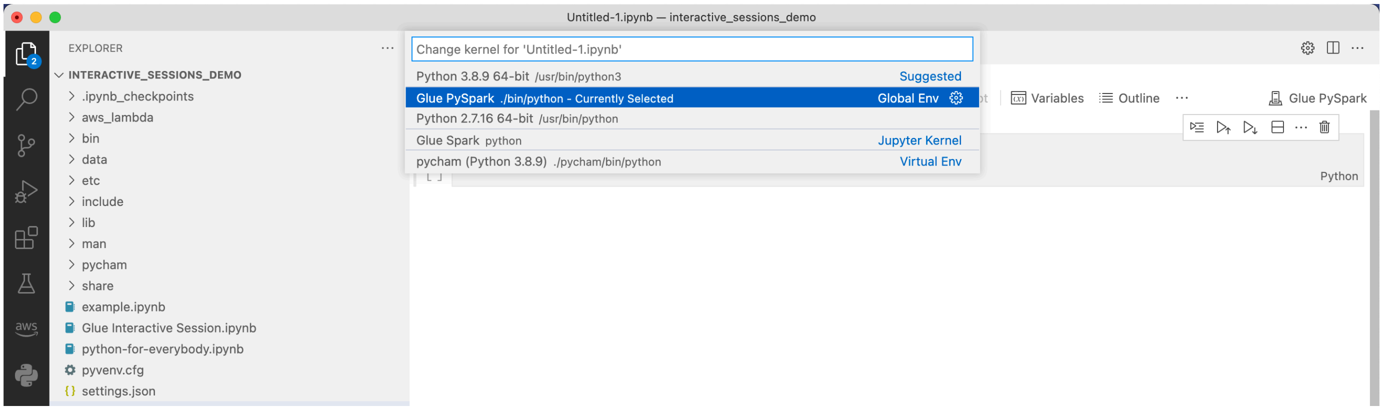
   2. In VS Code, click the current kernel, then **Select Another Kernel...**, then select **Existing Jupyter Server...**. Paste the URL you copied from the step above.

      If you receive an error message, see the [VS Code Jupyter wiki](#).

   3. If successful, this will set the kernel to **Glue PySpark**.


INTERACTIVE_SESSIONS_DEMO
.ipynb_checkpoints
aws_lambda
bin
data
etc
include
lib
man
pycham
share
Glue Interactive Session.ipynb
python-for-everybody.ipynb
pyvenv.cfg
settings.json
Untitled-1.ipynb
Untitled-2.ipynb
Untitled-2.ipynb > empty cell
+ Code
+ Markdown
Run All
Clear Outputs of All Cells
Outline
Select Kernel
Python

Choose the **Glue PySpark** or **Glue Spark** kernel (for Python and Scala respectively).


Untitled-1.ipynb — interactive_sessions_demo
EXPLORER
INTERACTIVE_SESSIONS_DEMO
.ipynb_checkpoints
aws_lambda
bin
data
etc
include
lib
man
pycham
share
example.ipynb
Glue Interactive Session.ipynb
python-for-everybody.ipynb
pyvenv.cfg
settings.json
Change kernel for 'Untitled-1.ipynb'
Python 3.8.9 64-bit /usr/bin/python3 Suggested
Glue PySpark ./bin/python - Currently Selected Global Env
Python 2.7.16 64-bit /usr/bin/python
Glue Spark python Jupyter Kernel
pycham (Python 3.8.9) ./pycham/bin/python Virtual Env
Variables
Outline
Glue PySpark
Python

If you don't see **AWS Glue PySpark** and **AWS Glue Spark** kernels in the drop-down list, please ensure you have installed the AWS Glue kernel in the step above, or that your `python.defaultInterpreterPath` setting in Visual Studio Code is correct. For more information, see [python.defaultInterpreterPath setting description](#).

5. Create an AWS Glue Interactive Session. Proceed to create a session in the same manner as you did in Jupyter Notebook. Specify any magics at the top of your first cell and run a statement of code.

Interactive sessions with IAM

These sections describe security considerations for AWS Glue interactive sessions.

Topics

- IAM principals used with interactive sessions
- Setting up a client principal

- Setting up a runtime role
- Make your session private with TagOnCreate
- IAM policy considerations

IAM principals used with interactive sessions

You use two IAM principals used with AWS Glue interactive sessions.

- **Client principal**: The client principal (either a user or a role) authorizes API operations for interactive sessions from an AWS Glue client that's configured with the principal's identity-based credentials. For example, this could be an IAM role that you typically use to access the AWS Glue console. This could also be a role given to a user in IAM whose credentials are used for the AWS Command Line Interface, or an AWS Glue client used by the interactive sessions Jupyter kernel.
- **Runtime role**: The runtime role is an IAM role that the client principal passes to interactive sessions API operations. AWS Glue uses this role to run statements in your session. For example, this role could be the one used for running AWS Glue ETL jobs.

  For more information, see Setting up a runtime role.

Setting up a client principal

You must attach an identity policy to the client principal to allow it to call the interactive sessions API. This role must have `iam:PassRole` access to the execution role that you would pass to the interactive sessions API, such as `CreateSession`. For example, you can attach the **AWSGlueConsoleFullAccess** managed policy to an IAM role which allows users in your account with the policy attached to access all the sessions created in your account (such as runtime statement or cancel statement).

If you would like to protect your session and make it private only to certain IAM roles, such as ones associated with the user who created the session then you can use AWS Glue Interactive Session's Tag Based Authorization Control called TagOnCreate. For more information, see Make your session private with TagOnCreate on how an owner tag-based scoped down managed policy can make your session private with TagOnCreate. For more information on identity-based policies, see Identity-based policies for AWS Glue.

Setting up a runtime role

You must pass an IAM role to the CreateSession API operation in order to allow AWS Glue to assume and run statements in interactive sessions. The role should have the same IAM permissions as those required to run a typical AWS Glue job. For example, you can create a service role using the **AWSGlueServiceRole** policy that allows AWS Glue to call AWS services on your behalf. If you use the AWS Glue console, it will automatically create a service role on your behalf or use an existing one. You can also create your own IAM role and attach your own IAM policy to allow similar permissions.

If you would like to protect your session and make it private only to the user who created the session then you can use AWS Glue Interactive Session's Tag Based Authorization Control called TagOnCreate. For more information, see Make your session private with TagOnCreate on how an owner tag-based scoped down managed policy can make your session private with TagOnCreate. For more information on identity-based policies, see Identity-based policies for AWS Glue. If you are creating the execution role by yourself from the IAM console and you want to make your service private with TagOnCreate feature then follow the steps below.

1. Create an IAM role with role type set to `Glue`.
2. Attach this AWS Glue managed policy: *AwsGlueSessionUserRestrictedServiceRole*
3. Prefix the role name with the policy name *AwsGlueSessionUserRestrictedServiceRole*. For example, you can create a role with name *AwsGlueSessionUserRestrictedServiceRole-myrole* and attach AWS Glue managed policy *AwsGlueSessionUserRestrictedServiceRole*.
4. Attach a trust policy like following to allow AWS Glue to assume the role:

   JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {
        "Service": [
          "glue.amazonaws.com"
        ]
      },
      "Action": [
        "sts:AssumeRole"
```

```
            ]
          }
        ]
      }
```

For an interactive sessions Jupyter kernel, you can specify the `iam_role` key in your AWS Command Line Interface profile. For more information, see [Configuring sessions with ~/.aws/config](). If you're interacting with interactive sessions using an AWS Glue notebook, then you can pass the execution role in the `%iam_role` magic in the first cell that you run.

Make your session private with TagOnCreate

AWS Glue interactive sessions supports tagging and Tag Based Authorization (TBAC) for interactive sessions as a named resource. In addition to TBAC using TagResource and UntagResource APIs, AWS Glue interactive sessions supports the TagOnCreate feature to 'tag' a session with a given tag only during session creation with CreateSession operation. This also means those tags will be removed on DeleteSession, aka UntagOnDelete.

TagOnCreate offers a powerful security mechanism to make your session private to the creator of the session. For example, you can attach an IAM policy with "owner" RequestTag and value of ${aws:userId} to a client principal (such as an user) in order to allow creating a session only if an "owner" tag with matching value of the callers userId is provided as userId tag in CreateSession request. This policy allows AWS Glue interactive sessions to create a session resource and tag the session with the userId tag only during session creation time. In addition to it you can scope down the access (like running statements) to your session only to the creator (aka owner tag with value ${aws:userId}) of the session by attaching an IAM policy with "owner" ResourceTag to the execution role you passed in during CreateSession.

In order to make it easier for you to use TagOnCreate feature to make a session private to the session creator, AWS Glue provides specialized managed policies and service roles.

If you want to create a AWS Glue Interactive Session using an IAM AssumeRole principal (that is, using credential vended by assuming an IAM role) and you want to make the session private to the creator, then use policies similar to the **AWSGlueSessionUserRestrictedNotebookPolicy** and **AWSGlueSessionUserRestrictedNotebookServiceRole** respectively. These policies allow AWS Glue to use ${aws:PrincipalTag} to extract the owner tag value. This requires you to pass a userId tag with value ${aws:userId} as SessionTag in the assume role credential. See [ID session tags](). If you are using an Amazon EC2 instance with an instance profile vending the credential and you want to create a session or interact with the session from within the Amazon EC2 instance , then you would require to pass a userId tag with value ${aws:userId} as SessionTag in the assume role credential.

For example, If you are creating a session using an IAM AssumeRole principal credential and you want to make your service private with TagOnCreate feature then follow the steps below.

1. Create a runtime role yourself from the IAM console. Please attach this AWS Glue managed policy *AwsGlueSessionUserRestrictedNotebookServiceRole* and prefix the role name with the policy name *AwsGlueSessionUserRestrictedNotebookServiceRole*. For example, you can create a role with name *AwsGlueSessionUserRestrictedNotebookServiceRole-myrole* and attach AWS Glue managed policy *AwsGlueSessionUserRestrictedNotebookServiceRole*.
2. Attach a trust policy like below to allow AWS Glue to assume the above role.

   JSON

   ```
   {
     "Version": "2012-10-17",
     "Statement": [
       {
         "Effect": "Allow",
         "Principal": {
           "Service": [
             "glue.amazonaws.com"
           ]
         },
         "Action": [
           "sts:AssumeRole"
         ]
       }
     ]
   }
   ```
3. Create another role named with a prefix *AwsGlueSessionUserRestrictedNotebookPolicy* and attach the AWS Glue managed policy *AwsGlueSessionUserRestrictedNotebookPolicy* to make the session private. In addition to the managed policy please attach the following inline policy to allow iam:PassRole to the role you created in step 1.

   JSON

   ```
   {
   ```

```
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "iam:PassRole"
      ],
      "Resource": [
        "arn:aws:iam::*:role/
AwsGlueSessionUserRestrictedNotebookServiceRole*"
      ],
      "Condition": {
        "StringLike": {
          "iam:PassedToService": [
            "glue.amazonaws.com"
          ]
        }
      }
    }
  ]
}
```

4. Attach a trust policy like following to the above IAM AWS Glue to assume the role.

   JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {
        "Service": [
          "glue.amazonaws.com"
        ]
      },
      "Action": [
        "sts:AssumeRole",
        "sts:TagSession"
      ]
    }
  ]
}
```

> **Note**
>
> Optionally, you can use a single role (for example, notebook role) and attach both of the above managed policies *AwsGlueSessionUserRestrictedNotebookServiceRole* and *AwsGlueSessionUserRestrictedNotebookPolicy*. Also attach the additional inline policy to allow `iam:passrole` of your role to AWS Glue. And finally attach the above trust policy to allow `sts:AssumeRole` and `sts:TagSession`.

AWSGlueSessionUserRestrictedNotebookPolicy

The AWSGlueSessionUserRestrictedNotebookPolicy provides access to create a AWS Glue Interactive Session from a notebook only if a tag key "owner" and value matching the AWS user id of the principal (user or Role). For more information, see Where you can use policy variables . This policy is attached to the principal (User or role) that creates AWS Glue Interactive Session notebooks from AWS Glue Studio. This policy also permits sufficient access to the AWS Glue Studio notebook to interact with the AWS Glue Studio Interactive Session resources that are created with the "owner" tag value matching the AWS user ID of the principal. This policy denies permission to change or remove "owner" tag from a AWS Glue session resource after the session is created.

AWSGlueSessionUserRestrictedNotebookServiceRole

The **AWSGlueSessionUserRestrictedNotebookServiceRole** provides sufficient access to the AWS Glue Studio notebook to interact with the AWS Glue Interactive Session resources that are created with the "owner" tag value matching the AWS user ID of the principal (user or role) of the notebook creator. For more information, see Where you can use policy variables . This service-role policy is attached to the role that is passed as magic to a notebook or passed as execution role to the CreateSession API. This policy also permits to create a AWS Glue Interactive Session from a notebook only if a tag key "owner" and value matching the AWS user ID of the principal. This policy denies permission to change or remove "owner" tag from an AWS Glue session resource after the session is created. This policy also includes permissions for writing and reading from Amazon S3 buckets, writing CloudWatch logs, creating and deleting tags for Amazon EC2 resources used by AWS Glue.

Make your session private with user policies

You can attach the **AWSGlueSessionUserRestrictedPolicy** to IAM roles attached to each of the users in your account to restrict them from creating a session only with an owner tag with a value matching their own ${aws:userId}. Instead of using the **AWSGlueSessionUserRestrictedNotebookPolicy** and **AWSGlueSessionUserRestrictedNotebookServiceRole** you need to use policies similar to the **AWSGlueSessionUserRestrictedPolicy** and **AWSGlueSessionUserRestrictedServiceRole** respectively. For more information, see Using-identity based policies . This policy scopes down the access to a session only to the creator, the ${aws:userId} of the user who created the session with an owner tag bearing their own ${aws:userId}. If you have created the execution role yourself using the IAM console by following the steps in Setting up a runtime role, then in addition to attaching the **AwsGlueSessionUserRestrictedPolicy** managed policy, also attach the following inline policy to each of the users in your account to allow `iam:PassRole` for the execution role you created earlier.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "iam:PassRole"
      ],
      "Resource": [
        "arn:aws:iam::*:role/AwsGlueSessionUserRestrictedServiceRole*"
      ],
      "Condition": {
        "StringLike": {
          "iam:PassedToService": [
            "glue.amazonaws.com"
          ]
        }
      }
    }
  ]
}
```

AWSGlueSessionUserRestrictedPolicy

The **AWSGlueSessionUserRestrictedPolicy** provides access to create an AWS Glue Interactive Session using the CreateSession API only if a tag key "owner" and value matching their AWS user ID is provided. This identity policy is attached to the user that invokes the CreateSession API. This policy also permits to interact with the AWS Glue Interactive Session resources that were created with a "owner" tag and value matching their AWS user id. This policy denies permission to change or remove "owner" tag from a AWS Glue session resource after the session is created.

AWSGlueSessionUserRestrictedServiceRole

The **AWSGlueSessionUserRestrictedServiceRole** provides full access to all AWS Glue resources except for sessions and allows users to create and use only the interactive sessions that are associated with the user. This policy also includes other permissions needed by AWS Glue to manage Glue resources in other AWS services. The policy also allows adding tags to AWS Glue resources in other AWS services.

IAM policy considerations

Interactive sessions are IAM resources in AWS Glue. Because they are IAM resources, access and interaction to a session is governed by IAM policies. Based on the IAM policies attached to a client principal or execution role configured by an admin, a client principal (user or role) will be able to create new sessions and interact with its own sessions and other sessions.

If an admin has attached an IAM policy such as AWSGlueConsoleFullAccess or AWSGlueServiceRole that allows access to all AWS Glue resources in that account, a client principal will be able to collaborate with each other. For example, one user will be able to interact with sessions that are created by other users if policies allow this.

If you'd like to configure a policy tailored to your specific needs, see IAM documentation about configuring resources for a policy . For example, in order to isolate sessions that belong to an user, you can use the TagOnCreate feature supported by AWS Glue Interactive sessions. See Make your session private with TagOnCreate .

Interactive sessions supports limiting session creation based on certain VPC conditions. See Control policies that control settings using condition keys.

Configuring AWS Glue interactive sessions for Jupyter and AWS Glue Studio notebooks

Introduction to Jupyter Magics

Jupyter Magics are commands that can be run at the beginning of a cell or as a whole cell body. Magics start with `%` for line-magics and `%%` for cell-magics. Line-magics such as `%region` and `%connections` can be run with multiple magics in a cell, or with code included in the cell body like the following example.

```
%region us-east-2
%connections my_rds_connection
dy_f = glue_context.create_dynamic_frame.from_catalog(database='rds_tables',
 table_name='sales_table')
```

Cell magics must use the entire cell and can have the command span multiple lines. An example of `%%sql` is below.

```
%%sql
select * from rds_tables.sales_table
```

Magics supported by AWS Glue interactive sessions for Jupyter

The following are magics that you can use with AWS Glue interactive sessions for Jupyter notebooks.

Sessions magics

Name	Type	Description
`%help`	n/a	Return a list of descriptions and input types for all magic commands.
`%profile`	String	Specify a profile in your AWS configura tion to use as the credentials provider.

Name	Type	Description
`%region`	String	Specify the AWS Region; in which to initialize a session. Default from `~/.aws/configure`. Example: `%region us-west-1`
`%idle_timeout`	Int	The number of minutes of inactivity after which a session will timeout after a cell has been executed. The default idle timeout value for Spark ETL sessions is the default timeout, 2880 minutes (48 hours). For other session types, consult documentation for that session type. Example: `%idle_timeout 3000`
`%session_id`	n/a	Return the session ID for the running session.
`%session_id_prefix`	String	Define a string that will precede all session IDs in the format **[session_id_prefix]-[session_id].** If a session ID is not provided, a random UUID will be generated. This magic is not supported when you run a Jupyter Notebook in AWS Glue Studio. Example: `%session_id_prefix 001`
`%status`		Return the status of the current AWS Glue session including its duration, configuration and executing user / role.
`%stop_session`		Stop the current session.
`%list_sessions`		Lists all currently running sessions by name and ID.

Name	Type	Description
`%session_type`	String	Sets the session type to one of Streaming, ETL, or Ray. Example: `%session_type Streaming`
`%glue_version`	String	The version of AWS Glue to be used by this session. Example: `%glue_version 3.0`

Magics for selecting job types

Name	Type	Description
`%streaming`	String	Changes the session type to AWS Glue Streaming.
`%etl`	String	Changes the session type to AWS Glue ETL.
%glue_ray	String	Changes the session type to AWS Glue for Ray. See Magics supported by AWS Glue Ray interactive sessions.

AWS Glue for Spark config magics

The `%%configure` magic is a json-formatted dictionary consisting of all configuration parameters for a session. Each parameter can be specified here or through individual magics.

Name	Type	Description
`%%configure`	Dictionary	Specify a JSON-formatted dictionary consisting of all configuration parameters for a session. Each parameter can be

Name	Type	Description
		specified here or through individual magics. For a list of parameters and examples on how to use `%%configure` , see [%%configure cell magic arguments].
`%iam_role`	String	Specify an IAM role ARN to execute your session with. Default from ~/.aws/configure. Example: `%iam_role AWSGlueServiceRole`
`%number_of_workers`	Int	The number of workers of a defined worker_type that are allocated when a job runs. `worker_type` must be set too. The default `number_of_workers` is 5. Example: `%number_of_workers 2`
`%additional_python_modules`	List	Comma separated list of additional Python modules to include in your cluster (can be from PyPI or S3). Example: `%additional_python_modules pandas, numpy` .

Name	Type	Description
`%%tags`	String	Adds tags to a session. Specify the tags within curly brackets { }. Each tag name pair is enclosed in parentheses (" ") and separated by a comma (,). Use the `%status` magic to view tags associated with the session.

```
%%tags
{"billing":"Data-Platform",
 "team":"analytics"}
```

```
%status
```

```
Session ID: <sessionId>
 Status: READY
 Role: <example-role>
 CreatedOn: 2023-05-26 11:12:17.
056000-07:00
 GlueVersion: 3.0
 Job Type: glueetl
 Tags: {'owner':'example-owner',
 'team':'analytics', 'billing'
:'Data-Platform'}
 Worker Type: G.4X
 Number of Workers: 5
 Region: us-west-2
 Applying the following default
 arguments:
 --glue_kernel_version 0.38.0
 --enable-glue-datacatalog true
 Arguments Passed: ['--glue_
kernel_version: 0.38.0', '--
enable-glue-datacatalog: true']
```

Name	Type	Description
`%%assume_role`	Dictionary	Specify a json-formatted dictionary or an IAM role ARN string to create a session for cross-account access. Example with ARN:

```
%%assume_role
{
  'arn:aws:iam::XXXXXXXXXXXX:
role/AWSGlueServiceRole'
}
```

Example with credentials:

```
 %%assume_role
{{
    "aws_access_key_id" =
 "XXXXXXXXXXXX",
    "aws_secret_access_key" =
 "XXXXXXXXXXXX",
    "aws_session_token" =
 "XXXXXXXXXXXX"
}}
```

%%configure cell magic arguments

The `%%configure` magic is a json-formatted dictionary consisting of all configuration parameters for a session. Each parameter can be specified here or through individual magics. See below for examples for arguments supported by the `%%configure` cell magic. Use the `--` prefix for run arguments specified for the job. Example:

```
%%configure
{
   "--user-jars-first": "true",
   "--enable-glue-datacatalog": "false"
}
```

For more information on job parameters, see Job parameters.

Session Configuration

Parameter	Type	Description
`max_retries`	Int	The maximum number of times to retry this job if it fails. `%%configure { "max_retries": "0" }`
`max_concurrent_runs`	Int	The maximum number of concurrent runs allowed for a job. Example: `%%configure { "max_concurrent_runs": "3" }`

`max_retries`:

```
%%configure
{
  "max_retries": "0"
}
```

`max_concurrent_runs`:

```
%%configure
{
  "max_concurrent_runs": "3"
}
```

Session parameters

Parameter	Type	Description
`--enable-spark-ui`	Boolean	Enable Spark UI to monitor and debug AWS Glue ETL jobs.

`--enable-spark-ui`:

```
%%configure
{
  "--enable-spark-ui": "true"
}
```

Parameter	Type	Description
`--spark-event-logs-path`	String	Specifies an Amazon S3 path. When using the Spark UI monitoring feature. Example:
`--script_location`	String	Specifies the S3 path to a script that executes a job. Example:
`--SECURITY_CONFIGURATION`	String	The name of a AWS Glue security configuration Example:

`--spark-event-logs-path`

```
%%configure
{
  "--spark-event-logs-path":
 "s3://path/to/event/logs/"
}
```

`--script_location`

```
%%configure
{
  "script_location": "s3://new-folder-here"
}
```

`--SECURITY_CONFIGURATION`

```
%%configure
{
    "--security_configuration":   {
"encryption_type": "kms",
"kms_key_id": "YOUR_KMS_KEY_ARN"
}
}
```

Parameter	Type	Description
`--job-language`	String	The script programming language. Accepts a value of 'scala' or 'python'. Default is 'python'. Example:

```
%%configure
{
  "--job-language": "scala"
}
```

Parameter	Type	Description
`--class`	String	The Scala class that serves as the entry point for your Scala script. Default is null. Example:

```
%%configure
{
  "--class": "className"
}
```

Parameter	Type	Description
`--user-jars-first`	Boolean	Prioritizes the customer's extra JAR files in the classpath. Default is null. Example:

```
%%configure
{
  "--user-jars-first": "true"
}
```

Parameter	Type	Description
`--use-postgres-driver`	Boolean	Prioritizes the Postgres JDBC driver in the class path to avoid a conflict with the Amazon Redshift JDBC driver. Default is null. Example:

```
%%configure
{
  "--use-postgres-driver": "true"
}
```

Parameter	Type	Description
`--extra-files`	List(string)	The Amazon S3 paths to additional files, such as configuration files that AWS Glue copies to the working directory of your script before executing it. Example:

```
%%configure
{
  "--extra-files": "s3://path/to/
additional/files/"
}
```

Parameter	Type	Description
`--job-bookmark-option`	String	Controls the behavior of a job bookmark. Accepts a value of 'job-bookmark-enable', 'job-bookmark-disable' or 'job-bookmark-pause'. Default is 'job-bookmark-disable'. Example:

```
%%configure
{
  "--job-bookmark-option": "job-bookmark-enable"
}
```

Parameter	Type	Description
`--TempDir`	String	Specifies an Amazon S3 path to a bucket that can be used as a temporary directory for the job. Default is null. Example:

```
%%configure
{
  "--TempDir": "s3://path/to/temp/dir"
}
```

Parameter	Type	Description
`--enable-s3-parquet-optimized-committer`	Boolean	Enables the EMRFS Amazon S3-optimized committer for writing Parquet data into Amazon S3. Default is 'true'. Example:

```
%%configure
{
  "--enable-s3-parquet-optimi
zed-committer": "false"
}
```

Parameter	Type	Description
`--enable-rename-algorithm-v2`	Boolean	Sets the EMRFS rename algorithm version to version 2. Default is 'true'. Example:

```
%%configure
{
  "--enable-rename-algorithm-
v2": "true"
}
```

Parameter	Type	Description
`--enable-glue-datacatalog`	Boolean	Enables you to use the AWS Glue Data Catalog as an Apache Spark Hive metastore. Example:

```
%%configure
{
  "--enable-glue-datacatalog":
 "true"
}
```

Parameter	Type	Description
`--enable-metrics`	Boolean	Enables the collection of metrics for job profiling for job run. Default is 'false'. Example:
`--enable-continuous-cloudwatch-log`	Boolean	Enables real-time continuous logging for AWS Glue jobs. Default is 'false'. Example:
`--enable-continuous-log-filter`	Boolean	Specifies a standard filter or no filter when you create or edit a job enabled for continuous logging. Default is 'true'. Example:

`--enable-metrics`

```
%%configure
{
  "--enable-metrics": "true"
}
```

`--enable-continuous-cloudwatch-log`

```
%%configure
{
  "--enable-continuous-cloudw
atch-log": "true"
}
```

`--enable-continuous-log-filter`

```
%%configure
{
  "--enable-continuous-log-fi
lter": "true"
}
```

Parameter	Type	Description
`--continuous-log-stream-prefix`	String	Specifies a custom Amazon CloudWatch log stream prefix for a job enabled for continuous logging. Default is null. Example:

```
%%configure
{
  "--continuous-log-stream-pr
efix": "prefix"
}
```

Parameter	Type	Description
`--continuous-log-conversionPattern`	String	Specifies a custom conversion log pattern for a job enabled for continuous logging. Default is null. Example:

```
%%configure
{
  "--continuous-log-conversio
nPattern": "pattern"
}
```

<table>
<tr><th>Parameter</th><th>Type</th><th>Description</th></tr>
<tr><td><code>--conf</code></td><td>String</td><td>

Controls Spark config parameters. It is for advanced use cases. Use `--conf` before each parameter. Example:

```
%%configure
{
    "--conf": "spark.hadoop.hive
.metastore.glue.catalogid=1
23456789012 --conf hive.meta
store.client.factory.class=
com.amazonaws.glue.catalog.
metastore.AWSGlueDataCatalo
gHiveClientFactory --conf
 hive.metastore.schema.verif
ication=false"
}
```

</td></tr>
<tr><td>timeout</td><td>Int</td><td>

Determines the maximum amount of time that the Spark session should wait for a statement to complete before terminating it.

```
%%configure
{
  "timeout": "30"
}
```

</td></tr>
<tr><td>auto-scaling</td><td>Boolean</td><td>

Determines whether or not to use auto-scaling.

```
%%configure
{
  "--enable-auto-scaling": "true"
}
```

</td></tr>
</table>

Spark jobs (ETL & streaming) magics

Name	Type	Description
`%worker_type`	String	Standard, G.1X, G.2X, G.4X, G.8X, G.12X, G.16X, R.1X, R.2X, R.4X, or R.8X. `number_of_workers` must be set too. The default worker_type is G.1X.
`%connections`	List	Specify a comma-separated list of connections to use in the session. Example:

```
%connections my_rds_connection
                    dy_f =
 glue_context.create_dynamic
_frame.from_catalog(databas
e='rds_tables', table_nam
e='sales_table')
```

Name	Type	Description
`%extra_py_files`	List	Comma separated list of additional Python files from Amazon S3.
`%extra_jars`	List	Comma-separated list of additional jars to include in the cluster.
`%spark_conf`	String	Specify custom spark configurations for your session. For example, `%spark_conf spark.serializer=org.apache.spark.serializer.KryoSerializer` .

Magics for Ray jobs

Name	Type	Description
`%min_workers`	Int	The minimum number of workers that are allocated to a Ray job. Default: 1. Example: `%min_workers 2`
`%object_memory_head`	Int	The percentage of free memory on the instance head node after a warm start. Minimum: 0. Maximum: 100. Example: `%object_memory_head 100`
%object_memory_worker	Int	The percentage of free memory on the instance worker nodes after a warm start. Minimum: 0. Maximum: 100. Example: `%object_memory_worker 100`

Action magics

Name	Type	Description
`%%sql`	String	Run SQL code. All lines after the initial `%%sql` magic will be passed as part of the SQL code. Example: `%%sql select * from rds_tables.sales_table`
`%matplot`	Matplotlib figure	Visualize your data using the matplotlib library. Example:

Name	Type	Description
		(code below)

```
import matplotlib.pyplot as plt

# Set X-axis and Y-axis values
x = [5, 2, 8, 4, 9]
y = [10, 4, 8, 5, 2]

# Create a bar chart
plt.bar(x, y)

# Show the plot
%matplot plt
```

Name	Type	Description
`%plotly`	Plotly figure	Visualize your data using the plotly library. Example:

```
import plotly.express as px

#Create a graphical figure
fig = px.line(x=["a","b","c"],
 y=[1,3,2], title="sample
 figure")

#Show the figure
%plotly fig
```

Naming sessions

AWS Glue interactive sessions are AWS resources and require a name. Names should be unique for each session and may be restricted by your IAM administrators. For more information, see Interactive sessions with IAM. The Jupyter kernel automatically generates unique session names for you. However sessions can be named manually in two ways:

1. Using the AWS Command Line Interface config file located at `~.aws/config`. See Setting Up AWS Config with the AWS Command Line Interface.

2. Using the `%session_id_prefix` magics. See [Magics supported by AWS Glue interactive sessions for Jupyter](#).

A session name is generated as follows:

- When the prefix and session_id are provided: the session name will be {prefix}-{UUID}.
- When nothing is provided: the session name will be {UUID}.

Prefixing session names allows you to recognize your session when listing it in the AWS CLI or console.

Specifying an IAM role for interactive sessions

You must specify an AWS Identity and Access Management (IAM) role to use with AWS Glue ETL code that you run with interactive sessions.

The role requires the same IAM permissions as those required to run AWS Glue jobs. See Create an IAM role for AWS Glue for more information on creating a role for AWS Glue jobs and interactive sessions.

IAM roles can be specified in two ways:

- Using the AWS Command Line Interface config file located at `~.aws/config` (Recommended). For more information, see Configuring sessions with ~/.aws/config.

> **Note**
>
> When the `%profile` magic is used, the configuration for `glue_iam_role` of that profile is honored.

- Using the %iam_role magic. For more information, see Magics supported by AWS Glue interactive sessions for Jupyter.

Configuring sessions with named profiles

AWS Glue interactive sessions uses the same credentials as the AWS Command Line Interface or boto3, and interactive sessions honors and works with named profiles like the AWS CLI found in

`~/.aws/config` (Linux and MacOS) or `%USERPROFILE%\.aws\config` (Windows). For more information, see [Using named profiles](#).

Interactive sessions takes advantage of named profiles by allowing the AWS Glue Service Role and Session ID Prefix to be specified in a profile. To configure a profile role, add a line for the `iam_role` key and/or `session_id_prefix` to your named profile as shown below. The `session_id_prefix` does not require quotes. For example, if you want to add a `session_id_prefix`, enter the value of the `session_id_prefix=myprefix`.

```
[default]
region=us-east-1
aws_access_key_id=AKIAIOSFODNN7EXAMPLE
aws_secret_access_key=wJalrXUtnFEMI/K7MDENG/bPxRfiCYEXAMPLEKEY
glue_iam_role=arn:aws:iam::<AccountID>:role/<GlueServiceRole>
session_id_prefix=<prefix_for_session_names>

[user1]
region=eu-west-1
aws_access_key_id=AKIAI44QH8DHBEXAMPLE
aws_secret_access_key=je7MtGbClwBF/2Zp9Utk/h3yCo8nvbEXAMPLEKEY
glue_iam_role=arn:aws:iam::<AccountID>:role/<GlueServiceRoleUser1>
session_id_prefix=<prefix_for_session_names_for_user1>
```

If you have a custom method of generating credentials, you can also configure your profile to use the `credential_process` parameter in your `~/.aws/config` file. For example:

```
[profile developer]
region=us-east-1
credential_process = "/Users/Dave/generate_my_credentials.sh" --username helen
```

You can find more details about sourcing credentials through the `credential_process` parameter here: [Sourcing credentials with an external process](#).

If a region or `iam_role` are not set in the profile that you are using, you must specify them using the `%region` and `%iam_role` magics in the first cell that you run.

Converting a script or notebook into an AWS Glue job

There are two ways you can convert a script or notebook into an AWS Glue job:

- Use **nbconvert** to convert your Jupyter `.ipynb` notebook document file into a `.py` file. For more information, see nbconvert: Convert Notebooks to other formats.
- Upload the file to AWS Glue Studio Notebooks.
  - In the AWS Glue Studio console, choose **Jobs** from the navigation menu.
  - In the **Create job** section, choose **Jupyter Notebook**.
  - In the **Options** section, choose **Upload and edit an existing notebook**.
  - Select **Choose file** to upload an `.ipynb` file.

Working with streaming operations in AWS Glue interactive sessions

Switching streaming session type

Use the AWS Glue interactive sessions configuration magic, `%streaming`, to define the job you are running and initialize a streaming interactive session.

Sampling input stream for interactive development

One tool we have derived to help enhance the interactive experience in AWS Glue interactive sessions is the addition of a new method under `GlueContext` to obtain a snapshot of a stream in a static DynamicFrame. `GlueContext` allows you to inspect, interact and implement your workflow.

With the `GlueContext` class instance, you will be able to locate the method `getSampleStreamingDynamicFrame`. Required arguments for this method are:

- `dataFrame`: The Spark Streaming DataFrame
- `options`: See available options below

Available options include：

- **windowSize**: This is also called Microbatch Duration. This parameter will determine how long a streaming query will wait after previous batch was triggered. This parameter value must be smaller than `pollingTimeInMs`.
- **pollingTimeInMs**: The total length of time the method will run. It will fire off at least one micro batch to obtain sample records from the input stream.

- **recordPollingLimit**: This parameter helps you limit the total number of records you will poll from the stream.
- (Optional) You can also use `writeStreamFunction` to apply this custom function to every record sampling function. See below for examples in Scala and Python.

Scala

```
val sampleBatchFunction = (batchDF: DataFrame, batchId: Long) => {//Optional but
 you can replace your own forEachBatch function here}
val jsonString: String = s"""{"pollingTimeInMs": "10000", "windowSize": "5
 seconds"}"""
val dynFrame = glueContext.getSampleStreamingDynamicFrame(YOUR_STREAMING_DF,
 JsonOptions(jsonString), sampleBatchFunction)
dynFrame.show()
```

Python

```
def sample_batch_function(batch_df, batch_id):
       //Optional but you can replace your own forEachBatch function here
options = {
            "pollingTimeInMs": "10000",
            "windowSize": "5 seconds",
        }
glue_context.getSampleStreamingDynamicFrame(YOUR_STREAMING_DF, options,
 sample_batch_function)
```

Note

When the sampled `DynFrame` is empty, it could be caused by a few reasons:

- The Streaming source is set to "Latest" and no new data has been ingested during the sampling period.
- The polling time is not enough to process the records it ingested. Data won't show up unless the whole batch has been processed.

Running streaming applications in interactive sessions

In AWS Glue interactive sessions, you can run a the AWS Glue streaming application like how you would create a streaming application in the AWS Glue Console. Since interactive sessions is session-based, encountering exceptions in the runtime does not cause the session to stop. We now have the added benefit of developing your batch function iteratively. For example:

```
def batch_function(data_frame, batch_id):
    log.info(data_frame.count())
    invalid_method_call()
glueContext.forEachBatch(frame=streaming_df, batch_function = batch_function, options =
 {**})
```

In the example above, we included an invalid usage of a method and unlike regular AWS Glue jobs which will exit the entire application, the user's coding context and definitions are fully preserved and the session is still operational. There is no need to bootstrap a new cluster and rerun all the preceding transformation. This allows you to focus on quickly iterating your batch function implementations to obtain desirable outcomes.

It is important to note that Interactive Session evaluates each statement in a blocking manner so that the session will only execute one statement at a time. Since streaming queries are continuous and never ending, sessions with active streaming queries won't be able to handle any follow up statements unless they are interrupted. You can issue the interruption command directly from Jupyter Notebook and our kernel will handle the cancellation for you.

Take the following sequence of statements which are waiting for execution as an example:

```
Statement 1:
      val number = df.count()
      #Spark Action with deterministic result
      Result: 5

Statement 2:
      streamingQuery.start().awaitTermination()
      #Spark Streaming Query that will be executing continously
      Result: Constantly updated with each microbatch

Statement 3:
      val number2 = df.count()
```

```
    #This will not be executed as previous statement will be running indefinitely
```

Developing and testing AWS Glue job scripts locally

When you develop and test your AWS Glue for Spark job scripts, there are multiple available options:

- AWS Glue Studio console
  - Visual editor
  - Script editor
  - AWS Glue Studio notebook
- Interactive sessions
  - Jupyter notebook
- Docker image
  - Local development
  - Remote development

You can choose any of the above options based on your requirements.

If you prefer no code or less code experience, the AWS Glue Studio visual editor is a good choice.

If you prefer an interactive notebook experience, AWS Glue Studio notebook is a good choice. For more information, see Using Notebooks with AWS Glue Studio and AWS Glue. If you want to use your own local environment, interactive sessions is a good choice. For more information, see Using interactive sessions with AWS Glue.

If you prefer local/remote development experience, the Docker image is a good choice. This helps you to develop and test AWS Glue for Spark job scripts anywhere you prefer without incurring AWS Glue cost.

If you prefer local development without Docker, installing the AWS Glue ETL library directory locally is a good choice.

Developing using AWS Glue Studio

The AWS Glue Studio visual editor is a graphical interface that makes it easy to create, run, and monitor extract, transform, and load (ETL) jobs in AWS Glue. You can visually compose data transformation workflows and seamlessly run them on AWS Glue's Apache Spark-based serverless ETL engine. You can inspect the schema and data results in each step of the job. For more information, see the AWS Glue Studio User Guide.

Developing using interactive sessions

Interactive sessions allow you to build and test applications from the environment of your choice. For more information, see Using interactive sessions with AWS Glue.

Develop and test AWS Glue jobs locally using a Docker image

For a production-ready data platform, the development process and CI/CD pipeline for AWS Glue jobs is a key topic. You can flexibly develop and test AWS Glue jobs in a Docker container. AWS Glue hosts Docker images on Docker Hub to set up your development environment with additional utilities. You can use your preferred IDE, notebook, or REPL using AWS Glue ETL library. This topic describes how to develop and test AWS Glue version 5.0 jobs in a Docker container using a Docker image.

Available Docker images

The following Docker images are available for AWS Glue on Amazon ECR:.

- For AWS Glue version 5.0: `public.ecr.aws/glue/aws-glue-libs:5`
- For AWS Glue version 4.0: `public.ecr.aws/glue/aws-glue-libs:glue_libs_4.0.0_image_01`
- For AWS Glue version 3.0: `public.ecr.aws/glue/aws-glue-libs:glue_libs_3.0.0_image_01`
- For AWS Glue version 2.0: `public.ecr.aws/glue/aws-glue-libs:glue_libs_2.0.0_image_01`

> **Note**
>
> AWS Glue Docker images are compatible with both x86_64 and arm64.

In this example, we use `public.ecr.aws/glue/aws-glue-libs:5` and run the container on a local machine (Mac, Windows, or Linux). This container image has been tested for AWS Glue version 5.0 Spark jobs. The image contains the following:

- Amazon Linux 2023
- AWS Glue ETL Library
- Apache Spark 3.5.4
- Open table format libraries; Apache Iceberg 1.7.1, Apache Hudi 0.15.0, and Delta Lake 3.3.0
- AWS Glue Data Catalog Client
- Amazon Redshift connector for Apache Spark
- Amazon DynamoDB connector for Apache Hadoop

To set up your container, pull the image from ECR Public Gallery and then run the container. This topic demonstrates how to run your container with the following methods, depending on your requirements:

- `spark-submit`
- REPL shell `(pyspark)`
- `pytest`
- Visual Studio Code

Prerequisites

Before you start, make sure that Docker is installed and the Docker daemon is running. For installation instructions, see the Docker documentation for Mac or Linux. The machine running the Docker hosts the AWS Glue container. Also make sure that you have at least 7 GB of disk space for the image on the host running the Docker.

For more information about restrictions when developing AWS Glue code locally, see Local development restrictions .

Configuring AWS

To enable AWS API calls from the container, set up AWS credentials by following steps. In the following sections, we will use this AWS named profile.

1. Create an AWS named profile .

2. Open `cmd` on Windows or a terminal on Mac/Linux and run the following command in a terminal:

   ```
   PROFILE_NAME="<your_profile_name>"
   ```

In the following sections, we use this AWS named profile.

If you're running Docker on Windows, choose the Docker icon (right-click) and choose **Switch to Linux containers** before pulling the image.

Run the following command to pull the image from ECR Public:

```
docker pull public.ecr.aws/glue/aws-glue-libs:5
```

Run the container

You can now run a container using this image. You can choose any of following based on your requirements.

spark-submit

You can run an AWS Glue job script by running the `spark-submit` command on the container.

1. Write your script and save it as `sample.py` in the example below and save it under the `/local_path_to_workspace/src/` directory using the following commands:

   ```
   $ WORKSPACE_LOCATION=/local_path_to_workspace
   $ SCRIPT_FILE_NAME=sample.py
   $ mkdir -p ${WORKSPACE_LOCATION}/src
   $ vim ${WORKSPACE_LOCATION}/src/${SCRIPT_FILE_NAME}
   ```

2. These variables are used in the docker run command below. The sample code (sample.py) used in the spark-submit command below is included in the appendix at the end of this topic.

   Run the following command to execute the `spark-submit` command on the container to submit a new Spark application:

   ```
   $ docker run -it --rm \
       -v ~/.aws:/home
   ```

```
    /hadoop/.aws \
    -v $WORKSPACE_LOCATION:/home/hadoop/workspace/ \
    -e AWS_PROFILE=$PROFILE_NAME \
    --name glue5_spark_submit \
    public.ecr.aws/glue/aws-glue-libs:5 \
    spark-submit /home/hadoop/workspace/src/$SCRIPT_FILE_NAME
```

3. (Optionally) Configure `spark-submit` to match your environment. For example, you can pass your dependencies with the `--jars` configuration. For more information, consult Dynamically Loading Spark Properties in the Spark documentation.

REPL shell (Pyspark)

You can run REPL (`read-eval-print loops`) shell for interactive development. Run the following command to execute the PySpark command on the container to start the REPL shell:

```
$ docker run -it --rm \
    -v ~/.aws:/home/hadoop/.aws \
    -e AWS_PROFILE=$PROFILE_NAME \
    --name glue5_pyspark \
    public.ecr.aws/glue/aws-glue-libs:5 \
    pyspark
```

You will see the following output:

```
Python 3.11.6 (main, Jan  9 2025, 00:00:00) [GCC 11.4.1 20230605 (Red Hat 11.4.1-2)] on
 linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use
 setLogLevel(newLevel).
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.5.4-amzn-0
      /_/

Using Python version 3.11.6 (main, Jan  9 2025 00:00:00)
Spark context Web UI available at None
Spark context available as 'sc' (master = local[*], app id = local-1740643079929).
SparkSession available as 'spark'.
```

```
>>>
```

With this REPL shell, you can code and test interactively.

Pytest

For unit testing, you can use `pytest` for AWS Glue Spark job scripts. Run the following commands for preparation.

```
$ WORKSPACE_LOCATION=/local_path_to_workspace
$ SCRIPT_FILE_NAME=sample.py
$ UNIT_TEST_FILE_NAME=test_sample.py
$ mkdir -p ${WORKSPACE_LOCATION}/tests
$ vim ${WORKSPACE_LOCATION}/tests/${UNIT_TEST_FILE_NAME}
```

Run the following command to run `pytest` using `docker run`:

```
$ docker run -i --rm \
    -v ~/.aws:/home/hadoop/.aws \
    -v $WORKSPACE_LOCATION:/home/hadoop/workspace/ \
    --workdir /home/hadoop/workspace \
    -e AWS_PROFILE=$PROFILE_NAME \
    --name glue5_pytest \
    public.ecr.aws/glue/aws-glue-libs:5 \
    -c "python3 -m pytest --disable-warnings"
```

Once `pytest` finishes executing unit tests, your output will look something like this:

```
============================= test session starts ==============================
platform linux -- Python 3.11.6, pytest-8.3.4, pluggy-1.5.0
rootdir: /home/hadoop/workspace
plugins: integration-mark-0.2.0
collected 1 item

tests/test_sample.py .                                                   [100%]

======================== 1 passed, 1 warning in 34.28s =========================
```

Setting up the container to use Visual Studio Code

To set up the container with Visual Studio Code, complete the following steps:

1. Install Visual Studio Code.
2. Install Python.
3. Install Visual Studio Code Remote - Containers
4. Open the workspace folder in Visual Studio Code.
5. Press `Ctrl+Shift+P` (Windows/Linux) or `Cmd+Shift+P` (Mac).
6. Type `Preferences: Open Workspace Settings (JSON)`.
7. Press Enter.
8. Paste the following JSON and save it.

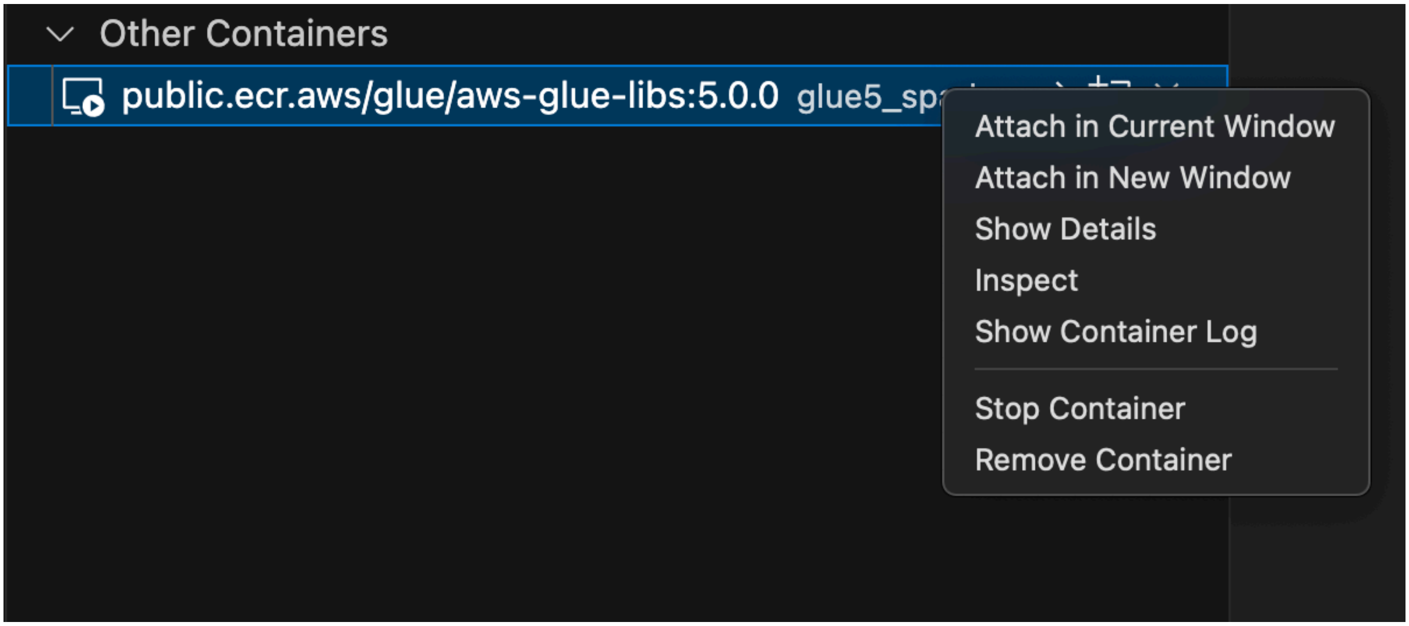
```
{
    "python.defaultInterpreterPath": "/usr/bin/python3.11",
    "python.analysis.extraPaths": [
        "/usr/lib/spark/python/lib/py4j-0.10.9.7-src.zip:/usr/lib/spark/python/:/usr/
lib/spark/python/lib/",
    ]
}
```

To set up the container:

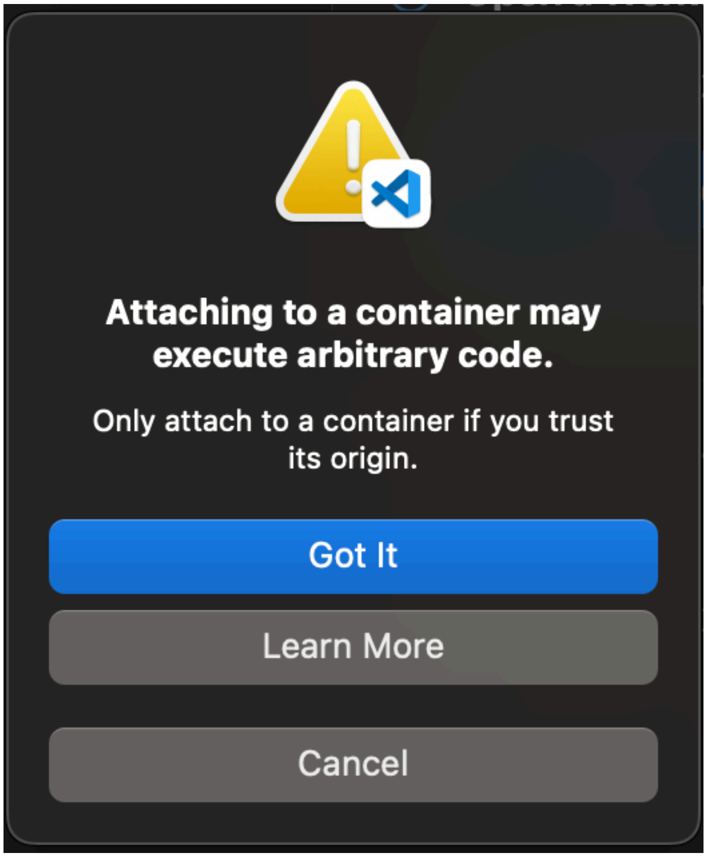
1. Run the Docker container.

```
$ docker run -it --rm \
    -v ~/.aws:/home/hadoop/.aws \
    -v $WORKSPACE_LOCATION:/home/hadoop/workspace/ \
    -e AWS_PROFILE=$PROFILE_NAME \
    --name glue5_pyspark \
    public.ecr.aws/glue/aws-glue-libs:5 \
    pyspark
```

2. Start Visual Studio Code.
3. Choose **Remote Explorer** on the left menu, and choose `amazon/aws-glue-libs:glue_libs_4.0.0_image_01`.
4. Right-click and choose **Attach in Current Window**.


Other Containers
public.ecr.aws/glue/aws-glue-libs:5.0.0 glue5_sp
Attach in Current Window
Attach in New Window
Show Details
Inspect
Show Container Log
Stop Container
Remove Container

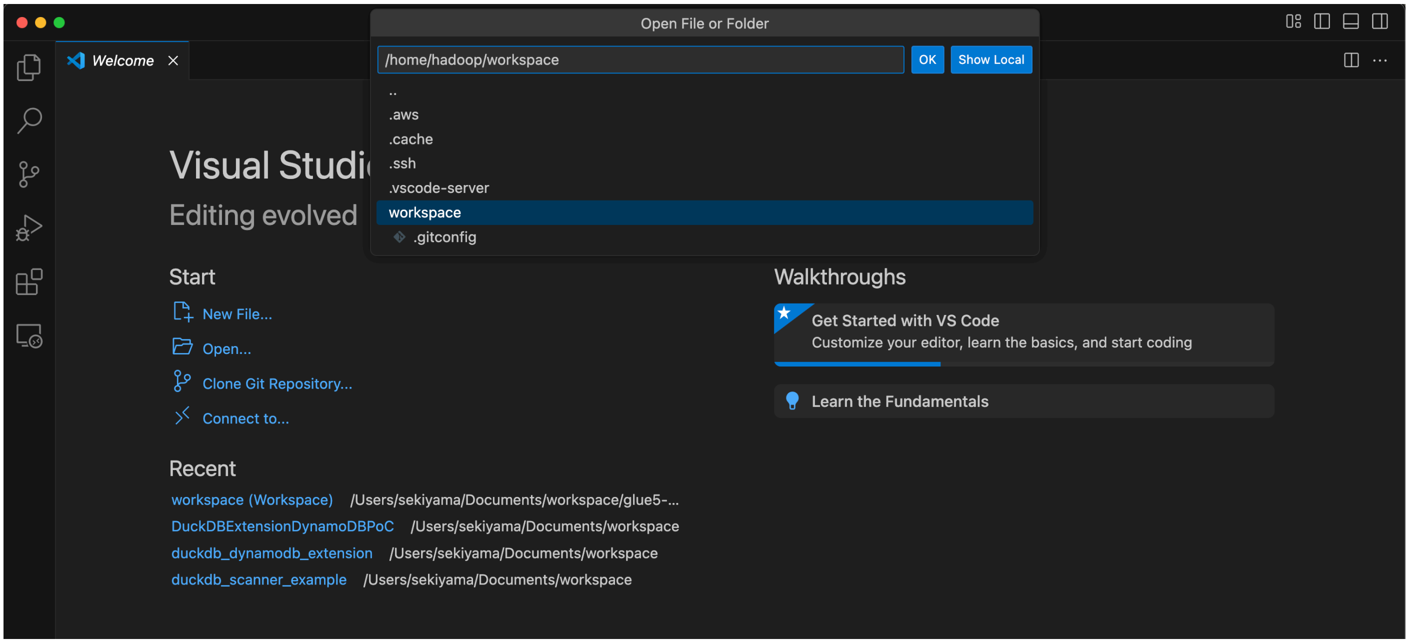
5. If the following dialog appears, choose **Got it**.


Attaching to a container may execute arbitrary code.

Only attach to a container if you trust its origin.

Got It
Learn More
Cancel

6. Open `/home/handoop/workspace/`.



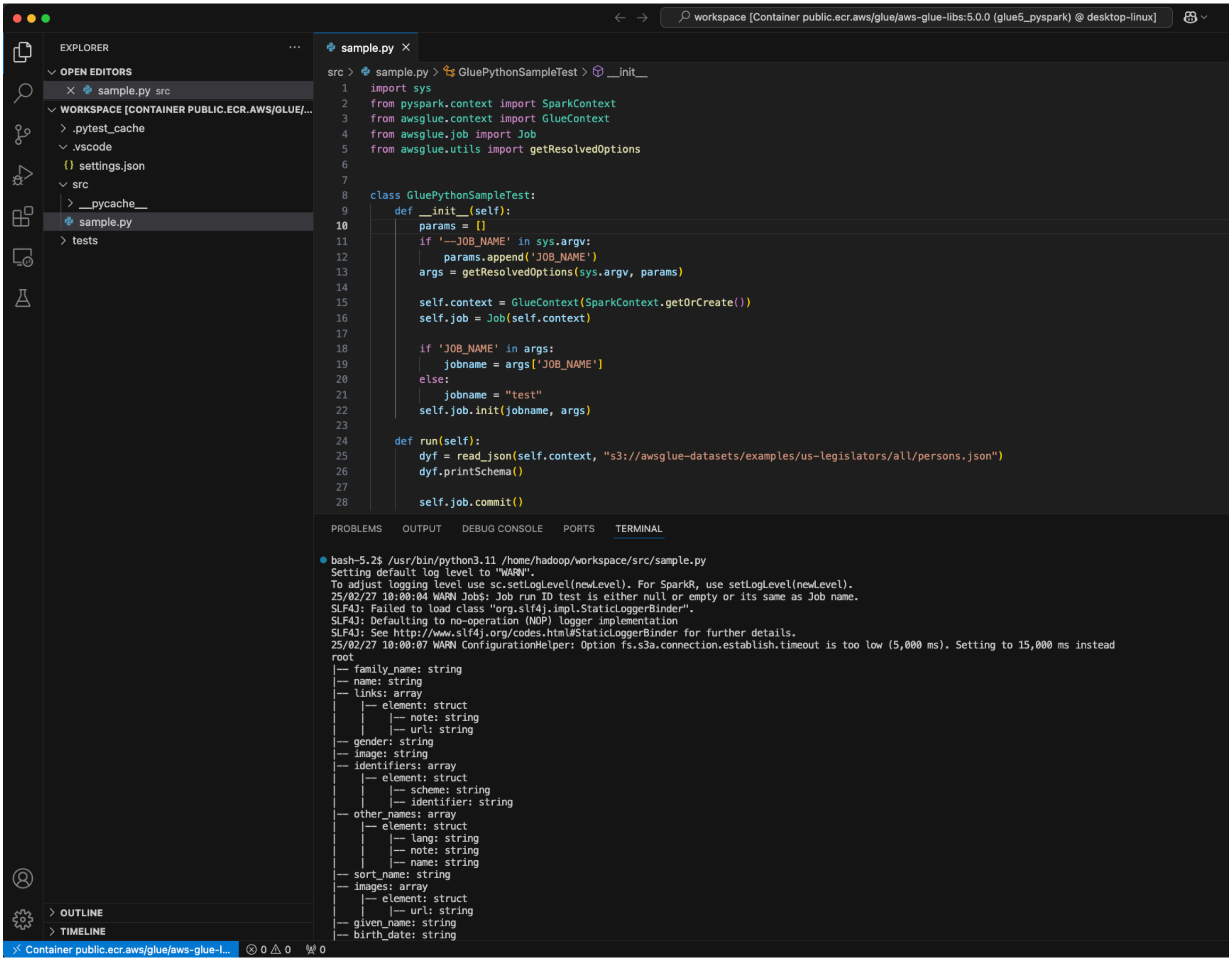
Open File or Folder

/home/hadoop/workspace

OK
Show Local
..
.aws
.cache
.ssh
.vscode-server
workspace
.gitconfig
Welcome
Visual Studi
Editing evolved
Start
New File...
Open...
Clone Git Repository...
Connect to...
Recent
workspace (Workspace) /Users/sekiyama/Documents/workspace/glue5-...
DuckDBExtensionDynamoDBPoC /Users/sekiyama/Documents/workspace
duckdb_dynamodb_extension /Users/sekiyama/Documents/workspace
duckdb_scanner_example /Users/sekiyama/Documents/workspace
Walkthroughs
Get Started with VS Code
Customize your editor, learn the basics, and start coding
Learn the Fundamentals

7. Create a AWS Glue PySpark script and choose **Run**.

   You will see the successful run of the script.


workspace [Container public.ecr.aws/glue/aws-glue-libs:5.0.0 (glue5_pyspark) @ desktop-linux]
EXPLORER
OPEN EDITORS
sample.py src
WORKSPACE [CONTAINER PUBLIC.ECR.AWS/GLUE/...
.pytest_cache
.vscode
settings.json
src
__pycache__
sample.py
tests
OUTLINE
TIMELINE
sample.py
src > sample.py > GluePythonSampleTest > __init__
PROBLEMS
OUTPUT
DEBUG CONSOLE
PORTS
TERMINAL
Container public.ecr.aws/glue/aws-glue-l...

```python
import sys
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job
from awsglue.utils import getResolvedOptions


class GluePythonSampleTest:
    def __init__(self):
        params = []
        if '--JOB_NAME' in sys.argv:
            params.append('JOB_NAME')
        args = getResolvedOptions(sys.argv, params)

        self.context = GlueContext(SparkContext.getOrCreate())
        self.job = Job(self.context)

        if 'JOB_NAME' in args:
            jobname = args['JOB_NAME']
        else:
            jobname = "test"
        self.job.init(jobname, args)

    def run(self):
        dyf = read_json(self.context, "s3://awsglue-datasets/examples/us-legislators/all/persons.json")
        dyf.printSchema()

        self.job.commit()
```

```
bash-5.2$ /usr/bin/python3.11 /home/hadoop/workspace/src/sample.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
25/02/27 10:00:04 WARN Job$: Job run ID test is either null or empty or its same as Job name.
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
25/02/27 10:00:07 WARN ConfigurationHelper: Option fs.s3a.connection.establish.timeout is too low (5,000 ms). Setting to 15,000 ms instead
root
|-- family_name: string
|-- name: string
|-- links: array
|    |-- element: struct
|    |    |-- note: string
|    |    |-- url: string
|-- gender: string
|-- image: string
|-- identifiers: array
|    |-- element: struct
|    |    |-- scheme: string
|    |    |-- identifier: string
|-- other_names: array
|    |-- element: struct
|    |    |-- lang: string
|    |    |-- note: string
|    |    |-- name: string
|-- sort_name: string
|-- images: array
|    |-- element: struct
|    |    |-- url: string
|-- given_name: string
|-- birth_date: string
```

Changes between AWS Glue 4.0 and AWS Glue 5.0 Docker image

The major changes between the AWS Glue 4.0 and AWS Glue 5.0 Docker image:

- In AWS Glue 5.0, there is a single container image for both batch and streaming jobs. This differs from Glue 4.0, where there was one image for batch and another for streaming.
- In AWS Glue 5.0, the default user name of the container is `hadoop`. In AWS Glue 4.0, the default user name was `glue_user`.
- In AWS Glue 5.0, several additional libraries including JupyterLab and Livy have been removed from the image. You can manually install them.
- In AWS Glue 5.0, all of Iceberg, Hudi and Delta libraries are pre-loaded by default, and the environment variable `DATALAKE_FORMATS` is no longer needed. Prior to AWS Glue 4.0, the

environment variable `DATALAKE_FORMATS` environment variable was used to specify which specific table formats should be loaded.

The above list is specific to the Docker image. To learn more about AWS Glue 5.0 updates, see Introducing AWS Glue 5.0 for Apache Spark and Migrating AWS Glue for Spark jobs to AWS Glue version 5.0.

Considerations

Keep in mind that the following features are not supported when using the AWS Glue container image to develop job scripts locally.

- Job bookmarks
- AWS Glue Parquet writer (Using the Parquet format in AWS Glue)
- FillMissingValues transform
- FindMatches transform
- Vectorized SIMD CSV reader
- The property customJdbcDriverS3Path for loading JDBC driver from Amazon S3 path
- AWS Glue Data Quality
- Sensitive Data Detection
- AWS Lake Formation permission-based credential vending

Appendix: Adding JDBC drivers and Java libraries

To add JDBC driver not currently available in the container, you can create a new directory under your workspace with JAR files you need and mount the directory to `/opt/spark/jars/` in docker run command. JAR files found under `/opt/spark/jars/` within the container are automatically added to Spark Classpath and will be available for use during job run.

For example, use the following docker run command to add JDBC driver jars to PySpark REPL shell.

```
docker run -it --rm \
    -v ~/.aws:/home/hadoop/.aws \
    -v $WORKSPACE_LOCATION:/home/hadoop/workspace/ \
    -v $WORKSPACE_LOCATION/jars/:/opt/spark/jars/ \
    --workdir /home/hadoop/workspace \
    -e AWS_PROFILE=$PROFILE_NAME \
    --name glue5_jdbc \
    public.ecr.aws/glue/aws-glue-libs:5 \
    pyspark
```

As highlighted in **Considerations**, `customJdbcDriverS3Path` connection option cannot be used to import a custom JDBC driver from Amazon S3 in AWS Glue container images.

Development endpoints

> **Note**
>
> **The console experience for dev endpoints has been removed as of March 31, 2023.** Creating, updating, and monitoring dev endpoints is still available via the [Development endpoints API]() and [AWS Glue CLI]().

We strongly recommend migrating from dev endpoints to interactive sessions for the reasons listed below. For required actions on how to migrate from dev endpoints to interactive sessions, see [Migrating from dev endpoints to interactive sessions]().

Description	Dev endpoints	Interactive sessions
Glue version support	Supports AWS Glue version 0.9 and 1.0	Supports AWS Glue version 2.0 and later
Dev endpoints are not available in the Asia Pacific (Jakarta) (`ap-southeast-3`), Middle East (UAE) (`me-central-1`), Europe (Spain) (`eu-south-2`), Europe (Zurich) (`eu-central-2`), or other new regions going forward	Interactive sessions are not currently available in the Middle East (UAE) (`me-central-1`) region, but may be made available later	
Access method to the Spark cluster	Supports SSH, REPL shell, Jupyter notebook, IDE (e.g. PyCharm)	supports AWS Glue Studio notebook, Jupyter notebook, various IDEs (for example,

Description	Dev endpoints	Interactive sessions
		Visual Studio Code, PyCharm), and SageMaker AI notebook
Time to first query	Requires 10-15 minutes to setup a Spark cluster	Can take up to 1 minute to set up an ephemeral Spark cluster
Price model	AWS charges for development endpoints based on the time that the endpoint is provisioned and the number of DPUs. Development endpoints do not time out. There is a 10-minute minimum billing duration for each provisioned development endpoint. Additionally, AWS charges for Jupyter notebook on Amazon EC2 instances, and SageMaker AI notebooks when you configure them with dev endpoints.	AWS charges for interactive sessions based on the time that the session is active and the number of DPUs. interactive sessions have configurable idle timeouts. AWS Glue Studio notebooks provide a built-in interface for interactive sessions and are offered at no additional cost. There is a 1-minute minimum billing duration for each interactive session. AWS Glue Studio notebooks provide a built-in interface for interactive sessions and are offered at no additional cost
Console experience	Only available via the CLI and API	Available through the AWS Glue console, CLI, and APIs

Migrating from dev endpoints to interactive sessions

Use the following checklist to determine the appropriate method to migrate from dev endpoints to interactive sessions.

Does your script depend on AWS Glue 0.9 or 1.0 specific features (for example, HDFS, YARN, etc.)?

If the answer is yes, see Migrating AWS Glue jobs to AWS Glue version 3.0. to learn how to migrate from Glue 0.9 or 1.0 to Glue 3.0 and later.

Which method do you use to access your dev endpoint?

If you use this method	Then do this
SageMaker AI notebook, Jupyter notebook, or JupyterLab	Migrate to AWS Glue Studio notebook by downloading `.ipynb` files on Jupyter and create a new AWS Glue Studio notebook job by uploading the `.ipynb` file. Alternatively, you can also use SageMaker AI Studio and select the AWS Glue kernel.
Zeppelin notebook	Convert the notebook to a Jupyter notebook manually by copying and pasting code or automatically using a third-party converter such as ze2nb. Then, use the notebook in AWS Glue Studio notebook or SageMaker AI Studio.
IDE	See Author AWS Glue jobs with PyCharm using AWS Glue interactive sessions, or Using interactive sessions with Microsoft Visual Studio Code.
REPL	Install the `aws-glue-session package` locally, then run the following command: • For Python: `jupyter console --kernal glue_pyspark` • For Scala: `jupyter console --kernal glue_spark`
SSH	No corresponding option on interactive sessions. Alternatively, you can use a Docker image. To learn more, see Developing using a Docker image.

The following sections provide information on using dev endpoints to develop jobs in AWS Glue version 1.0.

Topics

- Developing scripts using development endpoints
- Managing notebooks

Developing scripts using development endpoints

> **Note**
>
> Development Endpoints are only supported for versions of AWS Glue prior to 2.0. For an interactive environment where you can author and test ETL scripts, use Notebooks on AWS Glue Studio.

AWS Glue can create an environment—known as a *development endpoint*—that you can use to iteratively develop and test your extract, transform, and load (ETL) scripts. You can create, edit, and delete development endpoints using the AWS Glue console or API.

Managing your development environment

When you create a development endpoint, you provide configuration values to provision the development environment. These values tell AWS Glue how to set up the network so that you can access the endpoint securely and the endpoint can access your data stores.

You can then create a notebook that connects to the endpoint, and use your notebook to author and test your ETL script. When you're satisfied with the results of your development process, you can create an ETL job that runs your script. With this process, you can add functions and debug your scripts in an interactive manner.

Follow the tutorials in this section to learn how to use your development endpoint with notebooks.

Topics

- Development endpoint workflow
- How AWS Glue development endpoints work with SageMaker notebooks
- Adding a development endpoint
- Accessing your development endpoint

- Tutorial: Set up a Jupyter notebook in JupyterLab to test and debug ETL scripts
- Tutorial: Use a SageMaker AI notebook with your development endpoint
- Tutorial: Use a REPL shell with your development endpoint
- Tutorial: Set up PyCharm professional with a development endpoint
- Advanced configuration: sharing development endpoints among multiple users

Development endpoint workflow

To use an AWS Glue development endpoint, you can follow this workflow:

1. Create a development endpoint using the API. The endpoint is launched in a virtual private cloud (VPC) with your defined security groups.
2. The API polls the development endpoint until it is provisioned and ready for work. When it's ready, connect to the development endpoint using one of the following methods to create and test AWS Glue scripts.
    - Create an SageMaker AI notebook in your account. For more information about how to create a notebook, see the section called "Authoring code with AWS Glue Studio notebooks".
    - Open a terminal window to connect directly to a development endpoint.
    - If you have the professional edition of the JetBrains PyCharm Python IDE, connect it to a development endpoint and use it to develop interactively. If you insert `pydevd` statements in your script, PyCharm can support remote breakpoints.
3. When you finish debugging and testing on your development endpoint, you can delete it.

How AWS Glue development endpoints work with SageMaker notebooks

One of the common ways to access your development endpoints is to use Jupyter on SageMaker notebooks. The Jupyter notebook is an open-source web application which is widely used in visualization, analytics, machine learning, etc. An AWS Glue SageMaker notebook provides you a Jupyter notebook experience with AWS Glue development endpoints. In the AWS Glue SageMaker notebook, the Jupyter notebook environment is pre-configured with SparkMagic, an open source Jupyter plugin to submit Spark jobs to a remote Spark cluster. Apache Livy is a service that allows interaction with a remote Spark cluster over a REST API. In the AWS Glue SageMaker notebook, SparkMagic is configured to call the REST API against a Livy server running on an AWS Glue development endpoint.

The following text flow explains how each component works:

AWS Glue SageMaker notebook: (Jupyter # SparkMagic) # (network) # AWS Glue development endpoint: (Apache Livy # Apache Spark)

Once you run your Spark script written in each paragraph on a Jupyter notebook, the Spark code is submitted to the Livy server via SparkMagic, then a Spark job named "livy-session-N" runs on the Spark cluster. This job is called a Livy session. The Spark job will run while the notebook session is alive. The Spark job will be terminated when you shutdown the Jupyter kernel from the notebook, or when the session is timed out. One Spark job is launched per notebook (.ipynb) file.

You can use a single AWS Glue development endpoint with multiple SageMaker notebook instances. You can create multiple notebook files in each SageMaker notebook instance. When you open an each notebook file and run the paragraphs, then a Livy session is launched per notebook file on the Spark cluster via SparkMagic. Each Livy session corresponds to single Spark job.

Default behavior for AWS Glue development endpoints and SageMaker notebooks

The Spark jobs run based on the Spark configuration. There are multiple ways to set the Spark configuration (for example, Spark cluster configuration, SparkMagic's configuration, etc.).

By default, Spark allocates cluster resources to a Livy session based on the Spark cluster configuration. In the AWS Glue development endpoints, the cluster configuration depends on the worker type. Here's a table which explains the common configurations per worker type.

	Standard	G.1X	G.2X
`spark.driver.memory`	5G	10G	20G
`spark.executor.memory`	5G	10G	20G
`spark.executor.cores`	4	8	16

	Standard	G.1X	G.2X
`spark.dynamicAllocation.enabled`	TRUE	TRUE	TRUE

The maximum number of Spark executors is automatically calculated by combination of DPU (or `NumberOfWorkers`) and worker type.

	Standard	G.1X	G.2X
The number of max Spark executors	`(DPU - 1) * 2 - 1`	`(NumberOfWorkers - 1)`	`(NumberOfWorkers - 1)`

For example, if your development endpoint has 10 workers and the worker type is `G.1X`, then you will have 9 Spark executors and the entire cluster will have 90G of executor memory since each executor will have 10G of memory.

Regardless of the specified worker type, Spark dynamic resource allocation will be turned on. If a dataset is large enough, Spark may allocate all the executors to a single Livy session since `spark.dynamicAllocation.maxExecutors` is not set by default. This means that other Livy sessions on the same dev endpoint will wait to launch new executors. If the dataset is small, Spark will be able to allocate executors to multiple Livy sessions at the same time.

> **Note**
>
> For more information about how resources are allocated in different use cases and how you set a configuration to modify the behavior, see Advanced configuration: sharing development endpoints among multiple users.

Adding a development endpoint

Use development endpoints to iteratively develop and test your extract, transform, and load (ETL) scripts in AWS Glue. Working with development endpoints is only available through the AWS Command Line Interface.

1. In a command line window, enter a command similar to the following.

   ```
   aws glue create-dev-endpoint --endpoint-name "endpoint1" --role-arn
    "arn:aws:iam::account-id:role/role-name" --number-of-nodes "3" --glue-version
    "1.0" --arguments '{"GLUE_PYTHON_VERSION": "3"}' --region "region-name"
   ```

   This command specifies AWS Glue version 1.0. Because this version supports both Python 2 and Python 3, you can use the `arguments` parameter to indicate the desired Python version. If the `glue-version` parameter is omitted, AWS Glue version 0.9 is assumed. For more information about AWS Glue versions, see the Glue version job property.

   For information about additional command line parameters, see create-dev-endpoint in the *AWS CLI Command Reference*.
2. (Optional) Enter the following command to check the development endpoint status. When the status changes to `READY`, the development endpoint is ready to use.

   ```
   aws glue get-dev-endpoint --endpoint-name "endpoint1"
   ```

Accessing your development endpoint

When you create a development endpoint in a virtual private cloud (VPC), AWS Glue returns only a private IP address. The public IP address field is not populated. When you create a non-VPC development endpoint, AWS Glue returns only a public IP address.

If your development endpoint has a **Public address**, confirm that it is reachable with the SSH private key for the development endpoint, as in the following example.

```

ssh -i dev-endpoint-private-key.pem glue@public-address
```

Suppose that your development endpoint has a **Private address**, your VPC subnet is routable from the public internet, and its security groups allow inbound access from your client. In this case, follow these steps to attach an *Elastic IP address* to a development endpoint to allow access from the internet.

> **Note**
>
> If you want to use Elastic IP addresses, the subnet that is being used requires an internet gateway associated through the route table.

To access a development endpoint by attaching an Elastic IP address

1. Open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, choose **Dev endpoints**, and navigate to the development endpoint details page. Record the **Private address** for use in the next step.
3. Open the Amazon EC2 console at https://console.aws.amazon.com/ec2/.
4. In the navigation pane, under **Network & Security**, choose **Network Interfaces**.
5. Search for the **Private DNS (IPv4)** that corresponds to the **Private address** on the AWS Glue console development endpoint details page.

   You might need to modify which columns are displayed on your Amazon EC2 console. Note the **Network interface ID** (ENI) for this address (for example, `eni-12345678`).
6. On the Amazon EC2 console, under **Network & Security**, choose **Elastic IPs**.
7. Choose **Allocate new address**, and then choose **Allocate** to allocate a new Elastic IP address.
8. On the **Elastic IPs** page, choose the newly allocated **Elastic IP**. Then choose **Actions**, **Associate address**.
9. On the **Associate address** page, do the following:

   - For **Resource type**, choose **Network interface**.
   - In the **Network interface** box, enter the **Network interface ID** (ENI) for the private address.
   - Choose **Associate**.
10. Confirm that the newly associated Elastic IP address is reachable with the SSH private key that is associated with the development endpoint, as in the following example.

```
ssh -i dev-endpoint-private-key.pem glue@elastic-ip
```

For information about using a bastion host to get SSH access to the development endpoint's private address, see the AWS Security Blog post [Securely Connect to Linux Instances Running in a Private Amazon VPC](#).

Tutorial: Set up a Jupyter notebook in JupyterLab to test and debug ETL scripts

In this tutorial, you connect a Jupyter notebook in JupyterLab running on your local machine to a development endpoint. You do this so that you can interactively run, debug, and test AWS Glue extract, transform, and load (ETL) scripts before deploying them. This tutorial uses Secure Shell (SSH) port forwarding to connect your local machine to an AWS Glue development endpoint. For more information, see [Port forwarding](#) on Wikipedia.

Step 1: Install JupyterLab and Sparkmagic

You can install JupyterLab by using `conda` or `pip`. `conda` is an open-source package management system and environment management system that runs on Windows, macOS, and Linux. `pip` is the package installer for Python.

If you're installing on macOS, you must have Xcode installed before you can install Sparkmagic.

1. Install JupyterLab, Sparkmagic, and the related extensions.

```
$ conda install -c conda-forge jupyterlab
$ pip install sparkmagic
$ jupyter nbextension enable --py --sys-prefix widgetsnbextension
$ jupyter labextension install @jupyter-widgets/jupyterlab-manager
```

2. Check the `sparkmagic` directory from `Location`.

```
$ pip show sparkmagic | grep Location
Location: /Users/username/.pyenv/versions/anaconda3-5.3.1/lib/python3.7/site-
packages
```

3. Change your directory to the one returned for `Location`, and install the kernels for Scala and PySpark.

```
$ cd /Users/username/.pyenv/versions/anaconda3-5.3.1/lib/python3.7/site-packages
$ jupyter-kernelspec install sparkmagic/kernels/sparkkernel
$ jupyter-kernelspec install sparkmagic/kernels/pysparkkernel
```

4. Download a sample `config` file.

   ```
   $ curl -o ~/.sparkmagic/config.json https://raw.githubusercontent.com/jupyter-incubator/sparkmagic/master/sparkmagic/example_config.json
   ```

   In this configuration file, you can configure Spark-related parameters like `driverMemory` and `executorCores`.

Step 2: Start JupyterLab

When you start JupyterLab, your default web browser is automatically opened, and the URL `http://localhost:8888/lab/workspaces/{workspace_name}` is shown.

```
$ jupyter lab
```

Step 3: Initiate SSH port forwarding to connect to your development endpoint

Next, use SSH local port forwarding to forward a local port (here, `8998`) to the remote destination that is defined by AWS Glue (`169.254.76.1:8998`).

1. Open a separate terminal window that gives you access to SSH. In Microsoft Windows, you can use the BASH shell provided by Git for Windows, or you can install Cygwin.
2. Run the following SSH command, modified as follows:

   - Replace *`private-key-file-path`* with a path to the `.pem` file that contains the private key corresponding to the public key that you used to create your development endpoint.
   - If you're forwarding a different port than `8998`, replace `8998` with the port number that you're actually using locally. The address `169.254.76.1:8998` is the remote port and isn't changed by you.
   - Replace *`dev-endpoint-public-dns`* with the public DNS address of your development endpoint. To find this address, navigate to your development endpoint in the AWS Glue console, choose the name, and copy the **Public address** that's listed on the **Endpoint details** page.

   ```
   ssh -i private-key-file-path -NTL 8998:169.254.76.1:8998 glue@dev-endpoint-public-dns
   ```

You will likely see a warning message like the following:

```
The authenticity of host 'ec2-xx-xxx-xxx-xx.us-west-2.compute.amazonaws.com
 (xx.xxx.xxx.xx)'
can't be established.  ECDSA key fingerprint is SHA256:4e97875Brt+1wKzRko
+JflSnp21X7aTP3BcFnHYLEts.
Are you sure you want to continue connecting (yes/no)?
```

Enter **yes** and leave the terminal window open while you use JupyterLab.

3. Check that SSH port forwarding is working with the development endpoint correctly.

```
$ curl localhost:8998/sessions
{"from":0,"total":0,"sessions":[]}
```

Step 4: Run a simple script fragment in a notebook paragraph

Now your notebook in JupyterLab should work with your development endpoint. Enter the following script fragment into your notebook and run it.

1. Check that Spark is running successfully. The following command instructs Spark to calculate 1 and then print the value.

```
spark.sql("select 1").show()
```

2. Check if AWS Glue Data Catalog integration is working. The following command lists the tables in the Data Catalog.

```
spark.sql("show tables").show()
```

3. Check that a simple script fragment that uses AWS Glue libraries works.

   The following script uses the `persons_json` table metadata in the AWS Glue Data Catalog to create a `DynamicFrame` from your sample data. It then prints out the item count and the schema of this data.

```
import sys
from pyspark.context import SparkContext
from awsglue.context import GlueContext
```

```

# Create a Glue context
glueContext = GlueContext(SparkContext.getOrCreate())

# Create a DynamicFrame using the 'persons_json' table
persons_DyF = glueContext.create_dynamic_frame.from_catalog(database="legislators",
 table_name="persons_json")

# Print out information about *this* data
print("Count:  ", persons_DyF.count())
persons_DyF.printSchema()
```

The output of the script is as follows.

```

 Count:  1961
 root
 |-- family_name: string
 |-- name: string
 |-- links: array
 |    |-- element: struct
 |    |    |-- note: string
 |    |    |-- url: string
 |-- gender: string
 |-- image: string
 |-- identifiers: array
 |    |-- element: struct
 |    |    |-- scheme: string
 |    |    |-- identifier: string
 |-- other_names: array
 |    |-- element: struct
 |    |    |-- note: string
 |    |    |-- name: string
 |    |    |-- lang: string
 |-- sort_name: string
 |-- images: array
 |    |-- element: struct
 |    |    |-- url: string
 |-- given_name: string
 |-- birth_date: string
 |-- id: string
 |-- contact_details: array
 |    |-- element: struct
 |    |    |-- type: string
 |    |    |-- value: string
 |-- death_date: string
```

Troubleshooting

- During the installation of JupyterLab, if your computer is behind a corporate proxy or firewall, you might encounter HTTP and SSL errors due to custom security profiles managed by corporate IT departments.

  The following is an example of a typical error that occurs when `conda` can't connect to its own repositories:

  ```
  CondaHTTPError: HTTP 000 CONNECTION FAILED for url <https://repo.anaconda.com/pkgs/
  main/win-64/current_repodata.json>
  ```

  This might happen because your company can block connections to widely used repositories in Python and JavaScript communities. For more information, see Installation Problems on the JupyterLab website.
- If you encounter a *connection refused* error when trying to connect to your development endpoint, you might be using a development endpoint that is out of date. Try creating a new development endpoint and reconnecting.

Tutorial: Use a SageMaker AI notebook with your development endpoint

In AWS Glue, you can create a development endpoint and then create a SageMaker AI notebook to help develop your ETL and machine learning scripts. A SageMaker AI notebook is a fully managed machine learning compute instance running the Jupyter Notebook application.

1. In the AWS Glue console, choose **Dev endpoints** to navigate to the development endpoints list.
2. Select the check box next to the name of a development endpoint that you want to use, and on the **Action** menu, choose **Create SageMaker notebook**.
3. Fill out the **Create and configure a notebook** page as follows:
   a. Enter a notebook name.
   b. Under **Attach to development endpoint**, verify the development endpoint.

   c. Create or choose an AWS Identity and Access Management (IAM) role.

      Creating a role is recommended. If you use an existing role, ensure that it has the required permissions. For more information, see the section called “Step 6: Create an IAM policy for SageMaker AI notebooks”.

   d. (Optional) Choose a VPC, a subnet, and one or more security groups.

   e. (Optional) Choose an AWS Key Management Service encryption key.

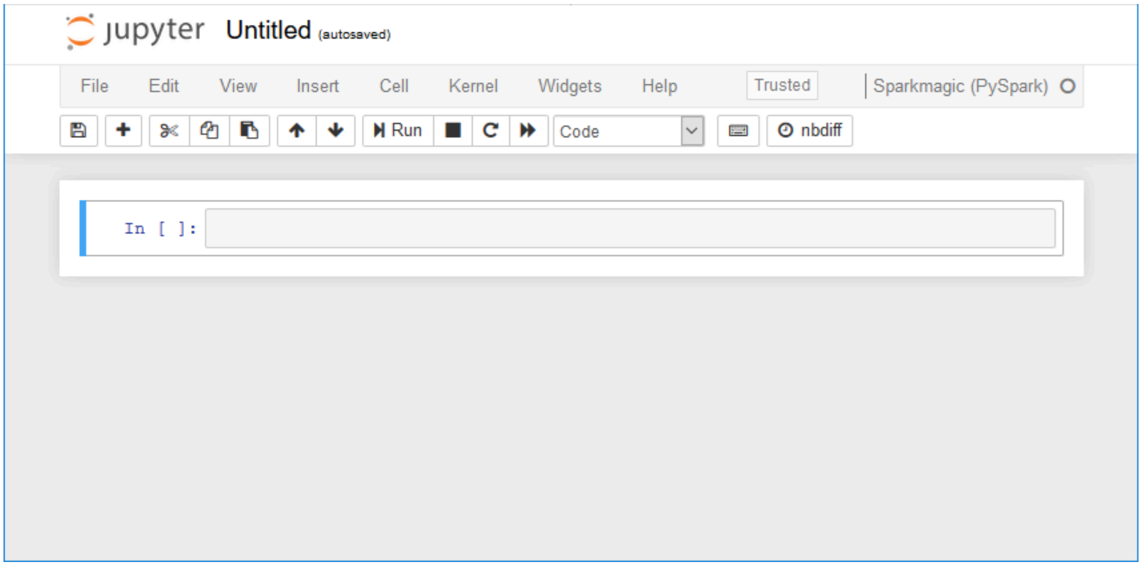
   f. (Optional) Add tags for the notebook instance.

4. Choose **Create notebook**. On the **Notebooks** page, choose the refresh icon at the upper right, and continue until the **Status** shows `Ready`.

5. Select the check box next to the new notebook name, and then choose **Open notebook**.

6. Create a new notebook: On the **jupyter** page, choose **New**, and then choose **Sparkmagic (PySpark)**.

   Your screen should now look like the following:

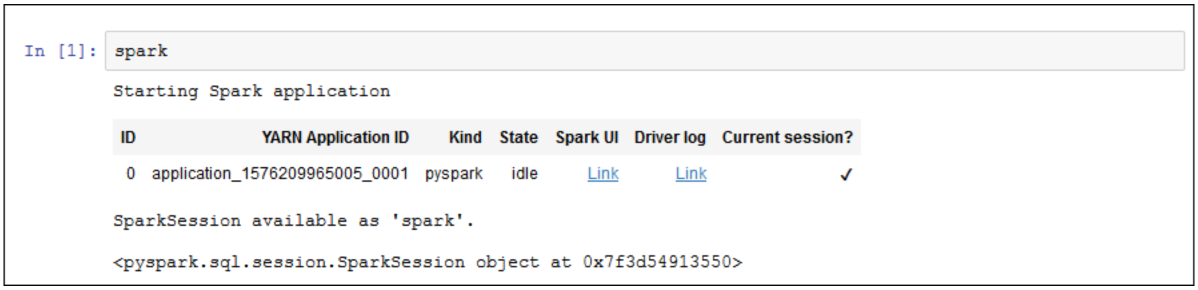


7. (Optional) At the top of the page, choose **Untitled**, and give the notebook a name.

8. To start a Spark application, enter the following command into the notebook, and then in the toolbar, choose **Run**.

```
spark
```

   After a short delay, you should see the following response:



```
In [1]: spark
Starting Spark application
```

ID	YARN Application ID	Kind	State	Spark UI	Driver log	Current session?
0	application_1576209965005_0001	pyspark	idle	Link	Link	✓

```
SparkSession available as 'spark'.
<pyspark.sql.session.SparkSession object at 0x7f3d54913550>
```

9. Create a dynamic frame and run a query against it: Copy, paste, and run the following code, which outputs the count and schema of the `persons_json` table.

```
import sys
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.transforms import *
glueContext = GlueContext(SparkContext.getOrCreate())
persons_DyF = glueContext.create_dynamic_frame.from_catalog(database="legislators",
 table_name="persons_json")
print ("Count:  ", persons_DyF.count())
persons_DyF.printSchema()
```

Tutorial: Use a REPL shell with your development endpoint

In AWS Glue, you can create a development endpoint and then invoke a REPL (Read–Evaluate–Print Loop) shell to run PySpark code incrementally so that you can interactively debug your ETL scripts before deploying them.

In order to use a REPL on a development endpoint, you need to have authorization to SSH to the endpoint.

1. On your local computer, open a terminal window that can run SSH commands, and paste in the edited SSH command. Run the command.

   Assuming that you accepted AWS Glue version 1.0 with Python 3 for the development endpoint, the output will look like this:

```
Python 3.6.8 (default, Aug  2 2019, 17:42:44)
[GCC 4.8.5 20150623 (Red Hat 4.8.5-28)] on linux
Type "help", "copyright", "credits" or "license" for more information.
SLF4J: Class path contains multiple SLF4J bindings.
```

```
SLF4J: Found binding in [jar:file:/usr/share/aws/glue/etl/jars/glue-assembly.jar!/
org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/spark/jars/slf4j-log4j12-1.7.16.jar!/
org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use
 setLogLevel(newLevel).
2019-09-23 22:12:23,071 WARN  [Thread-5] yarn.Client (Logging.scala:logWarning(66))
 - Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading
 libraries under SPARK_HOME.
2019-09-23 22:12:26,562 WARN  [Thread-5] yarn.Client (Logging.scala:logWarning(66))
 - Same name resource file:/usr/lib/spark/python/lib/pyspark.zip added multiple
 times to distributed cache
2019-09-23 22:12:26,580 WARN  [Thread-5] yarn.Client (Logging.scala:logWarning(66))
 - Same path resource file:///usr/share/aws/glue/etl/python/PyGlue.zip added
 multiple times to distributed cache.
2019-09-23 22:12:26,581 WARN  [Thread-5] yarn.Client (Logging.scala:logWarning(66))
 - Same path resource file:///usr/lib/spark/python/lib/py4j-src.zip added multiple
 times to distributed cache.
2019-09-23 22:12:26,581 WARN  [Thread-5] yarn.Client (Logging.scala:logWarning(66))
 - Same path resource file:///usr/share/aws/glue/libs/pyspark.zip added multiple
 times to distributed cache.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.4.3
      /_/

Using Python version 3.6.8 (default, Aug  2 2019 17:42:44)
SparkSession available as 'spark'.
>>>
```

2. Test that the REPL shell is working correctly by typing the statement, `print(spark.version)`. As long as that displays the Spark version, your REPL is now ready to use.

3. Now you can try executing the following simple script, line by line, in the shell:

```
import sys
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.transforms import *
glueContext = GlueContext(SparkContext.getOrCreate())
persons_DyF = glueContext.create_dynamic_frame.from_catalog(database="legislators",
 table_name="persons_json")
print ("Count:  ", persons_DyF.count())
persons_DyF.printSchema()
```

Tutorial: Set up PyCharm professional with a development endpoint

This tutorial shows you how to connect the PyCharm Professional Python IDE running on your local machine to a development endpoint so that you can interactively run, debug, and test AWS Glue ETL (extract, transfer, and load) scripts before deploying them. The instructions and screen captures in the tutorial are based on PyCharm Professional version 2019.3.

To connect to a development endpoint interactively, you must have PyCharm Professional installed. You can't do this using the free edition.

> **Note**
>
> The tutorial uses Amazon S3 as a data source. If you want to use a JDBC data source instead, you must run your development endpoint in a virtual private cloud (VPC). To connect with SSH to a development endpoint in a VPC, you must create an SSH tunnel. This tutorial does not include instructions for creating an SSH tunnel. For information on using SSH to connect to a development endpoint in a VPC, see Securely Connect to Linux Instances Running in a Private Amazon VPC in the AWS security blog.

Topics

- Connecting PyCharm professional to a development endpoint
- Deploying the script to your development endpoint
- Configuring a remote interpreter
- Running your script on the development endpoint

Connecting PyCharm professional to a development endpoint

1. Create a new pure-Python project in PyCharm named `legislators`.
2. Create a file named `get_person_schema.py` in the project with the following content:

```
from pyspark.context import SparkContext
from awsglue.context import GlueContext


def main():
    # Create a Glue context
    glueContext = GlueContext(SparkContext.getOrCreate())

    # Create a DynamicFrame using the 'persons_json' table
    persons_DyF =
 glueContext.create_dynamic_frame.from_catalog(database="legislators",
 table_name="persons_json")

    # Print out information about this data
    print("Count:  ", persons_DyF.count())
    persons_DyF.printSchema()


if __name__ == "__main__":
    main()
```

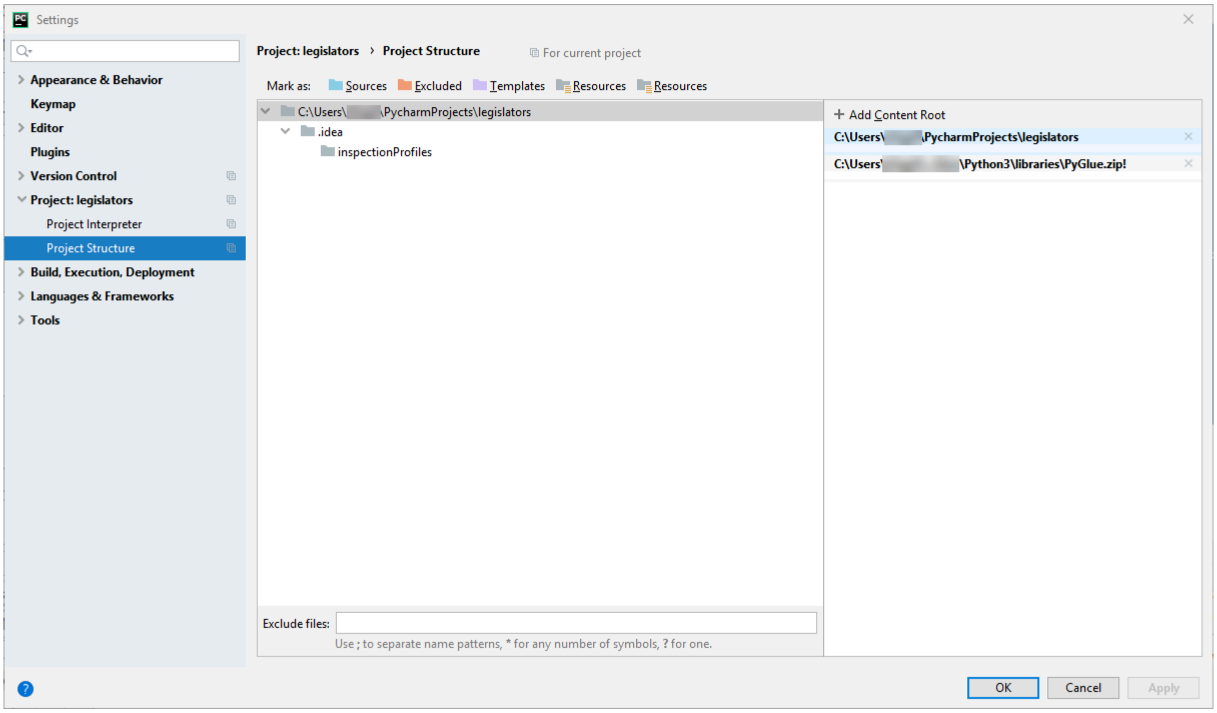
3. Do one of the following:

   - For AWS Glue version 0.9, download the AWS Glue Python library file, `PyGlue.zip`, from `https://s3.amazonaws.com/aws-glue-jes-prod-us-east-1-assets/etl/python/PyGlue.zip` to a convenient location on your local machine.
   - For AWS Glue version 1.0 and later, download the AWS Glue Python library file, `PyGlue.zip`, from `https://s3.amazonaws.com/aws-glue-jes-prod-us-east-1-assets/etl-1.0/python/PyGlue.zip` to a convenient location on your local machine.

4. Add `PyGlue.zip` as a content root for your project in PyCharm:

   - In PyCharm, choose **File**, **Settings** to open the **Settings** dialog box. (You can also press `Ctrl+Alt+S`.)
   - Expand the `legislators` project and choose **Project Structure**. Then in the right pane, choose **+ Add Content Root**.
   - Navigate to the location where you saved `PyGlue.zip`, select it, then choose **Apply**.

   The **Settings** screen should look something like the following:


Settings
Appearance & Behavior
Keymap
Editor
Plugins
Version Control
Project: legislators
Project Interpreter
Project Structure
Build, Execution, Deployment
Languages & Frameworks
Tools
Project: legislators › Project Structure
For current project
Mark as:
Sources
Excluded
Templates
Resources
Resources
+ Add Content Root
Exclude files:
Use ; to separate name patterns, * for any number of symbols, ? for one.
OK
Cancel
Apply

C:\Users\[redacted]\PycharmProjects\legislators
.idea
inspectionProfiles

C:\Users\[redacted]\PycharmProjects\legislators
C:\Users\[redacted]\Python3\libraries\PyGlue.zip!

Leave the **Settings** dialog box open after you choose **Apply**.

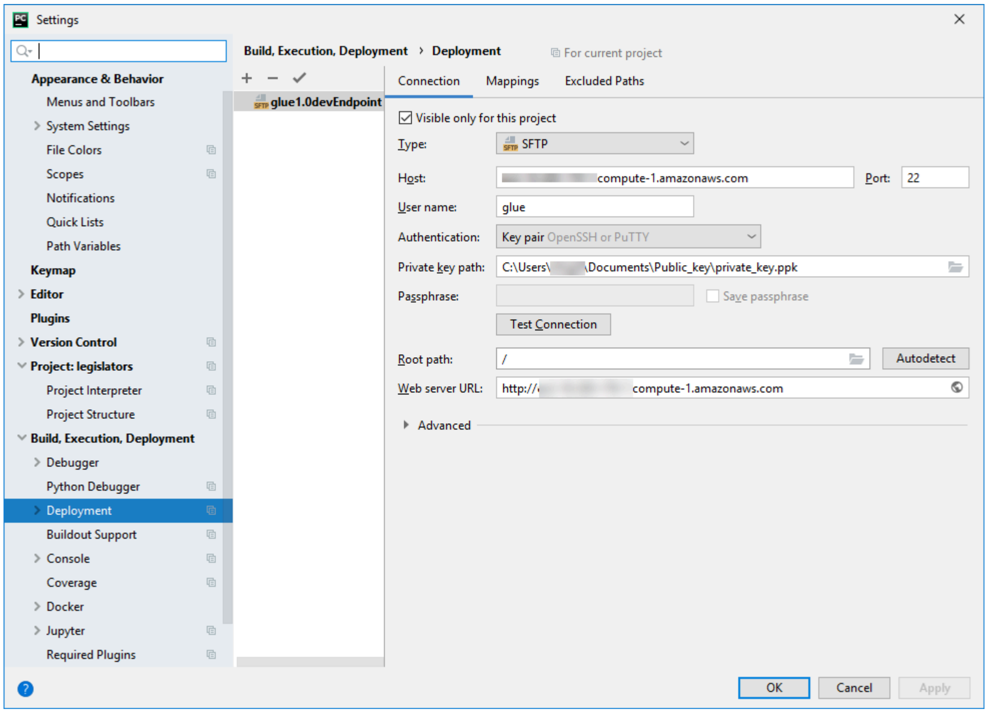
5. Configure deployment options to upload the local script to your development endpoint using SFTP (this capability is available only in PyCharm Professional):

   - In the **Settings** dialog box, expand the **Build, Execution, Deployment** section. Choose the **Deployment** subsection.
   - Choose the **+** icon at the top of the middle pane to add a new server. Set its **Type** to SFTP and give it a name.
   - Set the **SFTP host** to the **Public address** of your development endpoint, as listed on its details page. (Choose the name of your development endpoint in the AWS Glue console to display the details page). For a development endpoint running in a VPC, set **SFTP host** to the host address and local port of your SSH tunnel to the development endpoint.
   - Set the **User name** to `glue`.
   - Set the **Auth type** to **Key pair (OpenSSH or Putty)**. Set the **Private key file** by browsing to the location where your development endpoint's private key file is located. Note that PyCharm only supports DSA, RSA and ECDSA OpenSSH key types, and does not accept keys in Putty's private format. You can use an up-to-date version of `ssh-keygen` to generate a key-pair type that PyCharm accepts, using syntax like the following:

```
ssh-keygen -t rsa -f <key_file_name> -C "<your_email_address>"
```

- Choose **Test connection**, and allow the connection to be tested. If the connection succeeds, choose **Apply**.

The **Settings** screen should now look something like the following:


Settings
Build, Execution, Deployment › Deployment
For current project
Appearance & Behavior
Menus and Toolbars
System Settings
File Colors
Scopes
Notifications
Quick Lists
Path Variables
Keymap
Editor
Plugins
Version Control
Project: legislators
Project Interpreter
Project Structure
Build, Execution, Deployment
Debugger
Python Debugger
Deployment
Buildout Support
Console
Coverage
Docker
Jupyter
Required Plugins
glue1.0devEndpoint
Connection
Mappings
Excluded Paths
Test Connection
Autodetect
Advanced
OK
Cancel
Apply

☑ Visible only for this project

Type: SFTP

Host: [redacted]compute-1.amazonaws.com Port: 22

User name: glue

Authentication: Key pair OpenSSH or PuTTY

Private key path: C:\Users\[redacted]\Documents\Public_key\private_key.ppk

Passphrase: ☐ Save passphrase

Root path: /

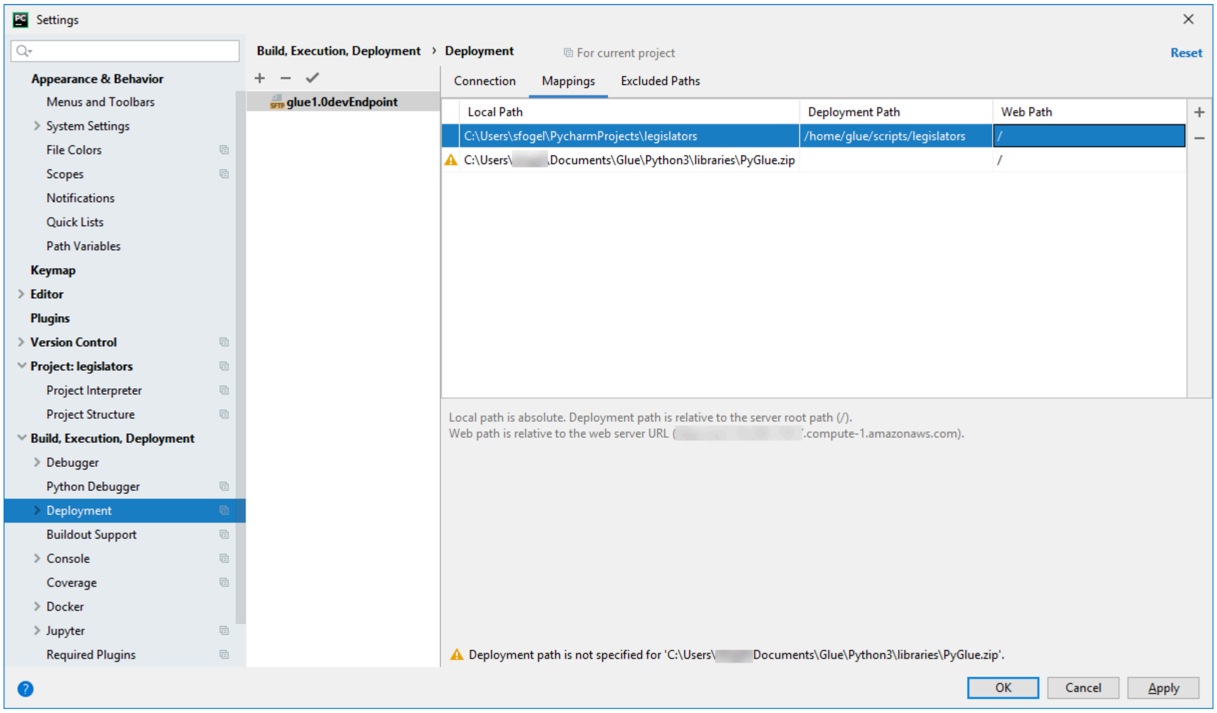
Web server URL: http://[redacted]compute-1.amazonaws.com

Again, leave the **Settings** dialog box open after you choose **Apply**.

6. Map the local directory to a remote directory for deployment:

   - In the right pane of the **Deployment** page, choose the middle tab at the top, labeled **Mappings**.
   - In the **Deployment Path** column, enter a path under `/home/glue/scripts/` for deployment of your project path. For example: `/home/glue/scripts/legislators`.
   - Choose **Apply**.

   The **Settings** screen should now look something like the following:


Settings
Appearance & Behavior
Menus and Toolbars
System Settings
File Colors
Scopes
Notifications
Quick Lists
Path Variables
Keymap
Editor
Plugins
Version Control
Project: legislators
Project Interpreter
Project Structure
Build, Execution, Deployment
Debugger
Python Debugger
Deployment
Buildout Support
Console
Coverage
Docker
Jupyter
Required Plugins
Build, Execution, Deployment › Deployment
For current project
Reset
glue1.0devEndpoint
Connection
Mappings
Excluded Paths
OK
Cancel
Apply

Local Path	Deployment Path	Web Path
C:\Users\sfogel\PycharmProjects\legislators	/home/glue/scripts/legislators	/
C:\Users\[redacted]Documents\Glue\Python3\libraries\PyGlue.zip		/

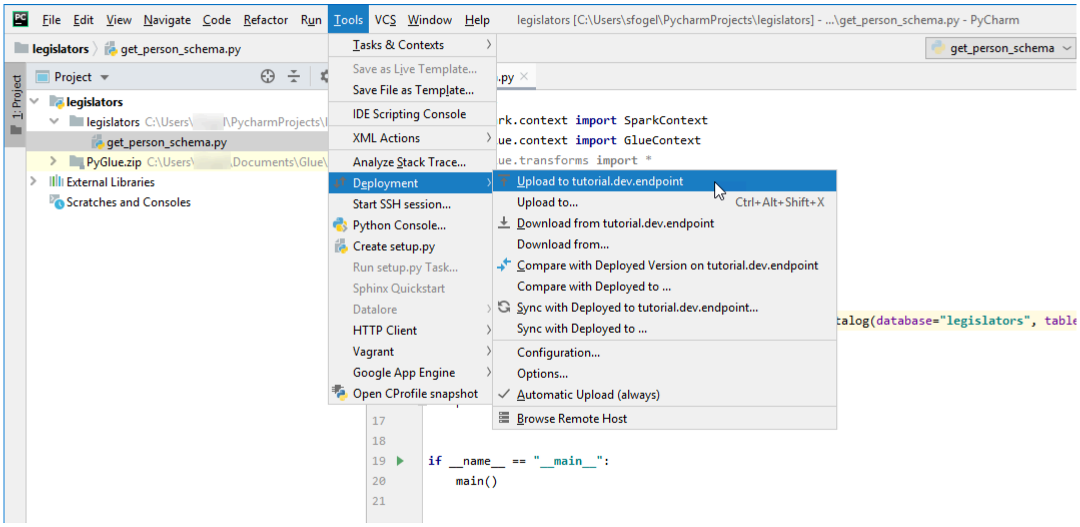
Local path is absolute. Deployment path is relative to the server root path (/).
Web path is relative to the web server URL ([redacted].compute-1.amazonaws.com).

Deployment path is not specified for 'C:\Users\[redacted]Documents\Glue\Python3\libraries\PyGlue.zip'.

Choose **OK** to close the **Settings** dialog box.

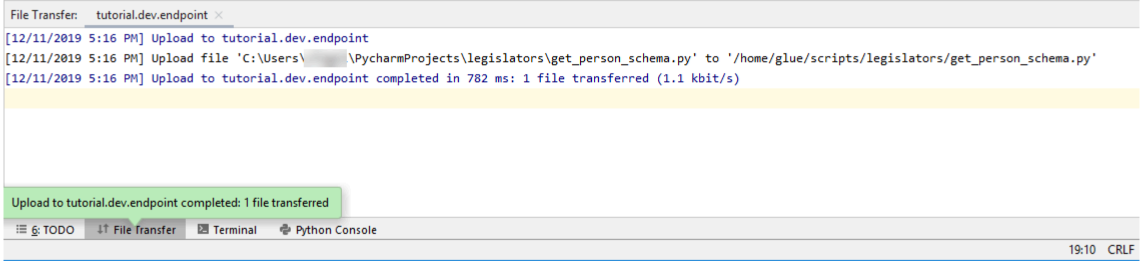
Deploying the script to your development endpoint

1. Choose **Tools**, **Deployment**, and then choose the name under which you set up your development endpoint, as shown in the following image:


File Edit View Navigate Code Refactor Run Tools VCS Window Help
legislators [C:\Users\sfogel\PycharmProjects\legislators] - ...\get_person_schema.py - PyCharm
legislators
get_person_schema.py
get_person_schema
Project
legislators
legislators C:\Users\[redacted]\PycharmProjects\l
get_person_schema.py
PyGlue.zip C:\Users\[redacted]Documents\Glue\
External Libraries
Scratches and Consoles
Tasks & Contexts
Save as Live Template...
Save File as Template...
IDE Scripting Console
XML Actions
Analyze Stack Trace...
Deployment
Start SSH session...
Python Console...
Create setup.py
Run setup.py Task...
Sphinx Quickstart
Datalore
HTTP Client
Vagrant
Google App Engine
Open CProfile snapshot
Upload to tutorial.dev.endpoint
Upload to... Ctrl+Alt+Shift+X
Download from tutorial.dev.endpoint
Download from...
Compare with Deployed Version on tutorial.dev.endpoint
Compare with Deployed to ...
Sync with Deployed to tutorial.dev.endpoint...
Sync with Deployed to ...
Configuration...
Options...
Automatic Upload (always)
Browse Remote Host

```
rk.context import SparkContext
ue.context import GlueContext
ue.transforms import *
talog(database="legislators", table
if __name__ == "__main__":
    main()
```

After your script has been deployed, the bottom of the screen should look something like the following:


File Transfer: tutorial.dev.endpoint
6: TODO
File Transfer
Terminal
Python Console
19:10 CRLF

```
[12/11/2019 5:16 PM] Upload to tutorial.dev.endpoint
[12/11/2019 5:16 PM] Upload file 'C:\Users\[redacted]\PycharmProjects\legislators\get_person_schema.py' to '/home/glue/scripts/legislators/get_person_schema.py'
[12/11/2019 5:16 PM] Upload to tutorial.dev.endpoint completed in 782 ms: 1 file transferred (1.1 kbit/s)
```

Upload to tutorial.dev.endpoint completed: 1 file transferred

2. On the menu bar, choose **Tools**, **Deployment**, **Automatic Upload (always)**. Ensure that a check mark appears next to **Automatic Upload (always)**.

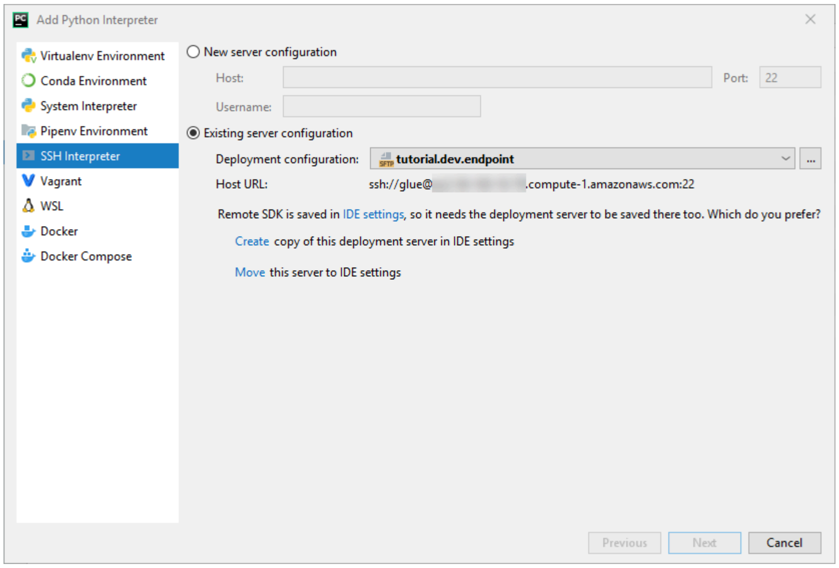
   When this option is enabled, PyCharm automatically uploads changed files to the development endpoint.

Configuring a remote interpreter

Configure PyCharm to use the Python interpreter on the development endpoint.

1. From the **File** menu, choose **Settings**.
2. Expand the project **legislators** and choose **Project Interpreter**.
3. Choose the gear icon next to the **Project Interpreter** list, and then choose **Add**.
4. In the **Add Python Interpreter** dialog box, in the left pane, choose **SSH Interpreter**.
5. Choose **Existing server configuration**, and in the **Deployment configuration** list, choose your configuration.

   Your screen should look something like the following image.


Add Python Interpreter
Virtualenv Environment
Conda Environment
System Interpreter
Pipenv Environment
SSH Interpreter
Vagrant
WSL
Docker
Docker Compose
Previous
Next
Cancel

○ New server configuration

Host:

Port: 22

Username:

● Existing server configuration

Deployment configuration: **tutorial.dev.endpoint**

Host URL: ssh://glue@[redacted].compute-1.amazonaws.com:22

Remote SDK is saved in IDE settings, so it needs the deployment server to be saved there too. Which do you prefer?

Create copy of this deployment server in IDE settings

Move this server to IDE settings

6. Choose **Move this server to IDE settings**, and then choose **Next**.
7. In the **Interpreter** field, change the path to `/usr/bin/gluepython` if you are using Python 2, or to `/usr/bin/gluepython3` if you are using Python 3. Then choose **Finish**.

Running your script on the development endpoint

To run the script:

- In the left pane, right-click the file name and choose **Run '*<filename>*'**.

  After a series of messages, the final output should show the count and the schema.

```
Count:   1961
root
|-- family_name: string
|-- name: string
|-- links: array
|    |-- element: struct
|    |    |-- note: string
|    |    |-- url: string
|-- gender: string
|-- image: string
|-- identifiers: array
|    |-- element: struct
|    |    |-- scheme: string
|    |    |-- identifier: string
|-- other_names: array
|    |-- element: struct
|    |    |-- lang: string
|    |    |-- note: string
|    |    |-- name: string
|-- sort_name: string
|-- images: array
|    |-- element: struct
|    |    |-- url: string
|-- given_name: string
|-- birth_date: string
|-- id: string
|-- contact_details: array
|    |-- element: struct
|    |    |-- type: string
|    |    |-- value: string
```

```
 |-- death_date: string


Process finished with exit code 0
```

You are now set up to debug your script remotely on your development endpoint.

Advanced configuration: sharing development endpoints among multiple users

This section explains how you can take advantage of development endpoints with SageMaker notebooks in typical use cases to share development endpoints among multiple users.

Single-tenancy configuration

In single tenant use-cases, to simplify the developer experience and to avoid contention for resources it is recommended that you have each developer use their own development endpoint sized for the project they are working on. This also simplifies the decisions related to worker type and DPU count leaving them up to the discretion of the developer and project they are working on.

You won't need to take care of resource allocation unless you runs multiple notebook files concurrently. If you run code in multiple notebook files at the same time, multiple Livy sessions will be launched concurrently. To segregate Spark cluster configurations in order to run multiple Livy sessions at the same time, you can follow the steps which are introduced in multi tenant use-cases.

For example, if your development endpoint has 10 workers and the worker type is `G.1X`, then you will have 9 Spark executors and the entire cluster will have 90G of executor memory since each executor will have 10G of memory.

Regardless of the specified worker type, Spark dynamic resource allocation will be turned on. If a dataset is large enough, Spark may allocate all the executors to a single Livy session since `spark.dynamicAllocation.maxExecutors` is not set by default. This means that other Livy sessions on the same dev endpoint will wait to launch new executors. If the dataset is small, Spark will be able to allocate executors to multiple Livy sessions at the same time.

Note

For more information about how resources are allocated in different use cases and how you set a configuration to modify the behavior, see Advanced configuration: sharing development endpoints among multiple users.

Multi-tenancy configuration

Note

Please note, development endpoints are intended to emulate the AWS Glue ETL environment as a single-tenant environment. While multi-tenant use is possible, it is an advanced use-case and it is recommended most users maintain a pattern of single-tenancy for each development endpoint.

In multi tenant use-cases, you might need to take care of resource allocation. The key factor is the number of concurrent users who use a Jupyter notebook at the same time. If your team works in a "follow-the-sun" workflow and there is only one Jupyter user at each time zone, then the number of concurrent users is only one, so you won't need to be concerned with resource allocation. However, if your notebook is shared among multiple users and each user submits code in an ad-hoc basis, then you will need to consider the below points.

To partition Spark cluster resources among multiple users, you can use SparkMagic configurations. There are two different ways to configure SparkMagic.

(A) Use the %%configure -f directive

If you want to modify the configuration per Livy session from the notebook, you can run the `%%configure -f` directive on the notebook paragraph.

For example, if you want to run Spark application on 5 executors, you can run the following command on the notebook paragraph.

```
%%configure -f
{"numExecutors":5}
```

Then you will see only 5 executors running for the job on the Spark UI.

We recommend limiting the maximum number of executors for dynamic resource allocation.

```
%%configure -f
{"conf":{"spark.dynamicAllocation.maxExecutors":"5"}}
```

(B) Modify the SparkMagic config file

SparkMagic works based on the Livy API. SparkMagic creates Livy sessions with configurations such as `driverMemory`, `driverCores`, `executorMemory`, `executorCores`, `numExecutors`, `conf`, etc. Those are the key factors that determine how much resources are consumed from the entire Spark cluster. SparkMagic allows you to provide a config file to specify those parameters which are sent to Livy. You can see a sample config file in this Github repository.

If you want to modify configuration across all the Livy sessions from a notebook, you can modify `/home/ec2-user/.sparkmagic/config.json` to add `session_config`.

To modify the config file on a SageMaker notebook instance, you can follow these steps.

1. Open a SageMaker notebook.
2. Open the Terminal kernel.
3. Run the following commands:

   ```
   sh-4.2$ cd .sparkmagic
   sh-4.2$ ls
   config.json logs
   sh-4.2$ sudo vim config.json
   ```

   For example, you can add these lines to `/home/ec2-user/.sparkmagic/config.json` and restart the Jupyter kernel from the notebook.

   ```
     "session_configs": {
       "conf": {
         "spark.dynamicAllocation.maxExecutors":"5"
       }
     },
   ```

Guidelines and best practices

To avoid this kind of resource conflict, you can use some basic approaches like:

- Have a larger Spark cluster by increasing the `NumberOfWorkers` (scaling horizontally) and upgrading the `workerType` (scaling vertically)
- Allocate fewer resources per user (fewer resources per Livy session)

Your approach will depend on your use case. If you have a larger development endpoint, and there is not a huge amount of data, the possibility of a resource conflict will decrease significantly because Spark can allocate resources based on a dynamic allocation strategy.

As described above, the number of Spark executors can be automatically calculated based on a combination of DPU (or `NumberOfWorkers`) and worker type. Each Spark application launches one driver and multiple executors. To calculate you will need the `NumberOfWorkers` = `NumberOfExecutors + 1`. The matrix below explains how much capacity you need in your development endpoint based on the number of concurrent users.

Number of concurrent notebook users	Number of Spark executors you want to allocate per user	Total NumberOfWorkers for your dev endpoint
3	5	18
10	5	60
50	5	300

If you want to allocate fewer resources per user, the `spark.dynamicAllocation.maxExecutors` (or `numExecutors`) would be the easiest parameter to configure as a Livy session parameter. If you set the below configuration in `/home/ec2-user/.sparkmagic/config.json`, then SparkMagic will assign a maximum of 5 executors per Livy session. This will help segregating resources per Livy session.

```
"session_configs": {
    "conf": {
      "spark.dynamicAllocation.maxExecutors":"5"
    }
  },
```

Suppose there is a dev endpoint with 18 workers (G.1X) and there are 3 concurrent notebook users at the same time. If your session config has `spark.dynamicAllocation.maxExecutors=5` then each user can make use of 1 driver and 5 executors. There won't be any resource conflicts even when you run multiple notebook paragraphs at the same time.

Trade-offs

With this session config `"spark.dynamicAllocation.maxExecutors":"5"`, you will be able to avoid resource conflict errors and you do not need to wait for resource allocation when there are concurrent user accesses. However, even when there are many free resources (for example, there are no other concurrent users), Spark cannot assign more than 5 executors for your Livy session.

Other notes

It is a good practice to stop the Jupyter kernel when you stop using a notebook. This will free resources and other notebook users can use those resources immediately without waiting for kernel expiration (auto-shutdown).

Common issues

Even when following the guidelines, you may experience certain issues.

Session not found

When you try to run a notebook paragraph even though your Livy session has been already terminated, you will see the below message. To activate the Livy session, you need to restart the Jupyter kernel by choosing **Kernel** > **Restart** in the Jupyter menu, then run the notebook paragraph again.

```
An error was encountered:
Invalid status code '404' from http://localhost:8998/sessions/13 with error payload:
 "Session '13' not found."
```

Not enough YARN resources

When you try to run a notebook paragraph even though your Spark cluster does not have enough resources to start a new Livy session, you will see the below message. You can often avoid this issue by following the guidelines, however, there might be a possibility that you face this issue. To workaround the issue, you can check if there are any unneeded, active Livy sessions. If there are unneeded Livy sessions, you will need to terminate them to free the cluster resources. See the next section for details.

```
Warning: The Spark session does not have enough YARN resources to start.
The code failed because of a fatal error:
    Session 16 did not start up in 60 seconds..
```

```
Some things to try:
a) Make sure Spark has enough available resources for Jupyter to create a Spark
 context.
b) Contact your Jupyter administrator to make sure the Spark magics library is
 configured correctly.
c) Restart the kernel.
```

Monitoring and debugging

This section describes techniques for monitoring resources and sessions.

Monitoring and debugging cluster resource allocation

You can watch the Spark UI to monitor how many resources are allocated per Livy session, and what are the effective Spark configurations on the job. To activate the Spark UI, see Enabling the Apache Spark Web UI for Development Endpoints.

(Optional) If you need a real-time view of the Spark UI, you can configure an SSH tunnel against the Spark history server running on the Spark cluster.

```
ssh -i <private-key.pem> -N -L 8157:<development endpoint public address>:18080
 glue@<development endpoint public address>
```

You can then open http://localhost:8157 on your browser to view the Spark UI.

Free unneeded Livy sessions

Review these procedures to shut down any unneeded Livy sessions from a notebook or a Spark cluster.

(a). Terminate Livy sessions from a notebook

You can shut down the kernel on a Jupyter notebook to terminate unneeded Livy sessions.

(b). Terminate Livy sessions from a Spark cluster

If there are unneeded Livy sessions which are still running, you can shut down the Livy sessions on the Spark cluster.

As a pre-requisite to perform this procedure, you need to configure your SSH public key for your development endpoint.

To log in to the Spark cluster, you can run the following command:

```
$ ssh -i <private-key.pem> glue@<development endpoint public address>
```

You can run the following command to see the active Livy sessions:

```
$ yarn application -list
20/09/25 06:22:21 INFO client.RMProxy: Connecting to ResourceManager at
 ip-255-1-106-206.ec2.internal/172.38.106.206:8032
Total number of applications (application-types: [] and states: [SUBMITTED, ACCEPTED,
 RUNNING]):2
Application-Id Application-Name Application-Type User Queue State Final-State Progress
 Tracking-URL
application_1601003432160_0005 livy-session-4 SPARK livy default RUNNING UNDEFINED 10%
 http://ip-255-1-4-130.ec2.internal:41867
application_1601003432160_0004 livy-session-3 SPARK livy default RUNNING UNDEFINED 10%
 http://ip-255-1-179-185.ec2.internal:33727
```

You can then shut down the Livy session with the following command:

```
$ yarn application -kill application_1601003432160_0005
20/09/25 06:23:38 INFO client.RMProxy: Connecting to ResourceManager at
 ip-255-1-106-206.ec2.internal/255.1.106.206:8032
Killing application application_1601003432160_0005
20/09/25 06:23:39 INFO impl.YarnClientImpl: Killed application
 application_1601003432160_0005
```

Managing notebooks

> **Note**
>
> Development Endpoints are only supported for versions of AWS Glue prior to 2.0. For an interactive environment where you can author and test ETL scripts, use Notebooks on AWS Glue Studio.

A notebook enables interactive development and testing of your ETL (extract, transform, and load) scripts on a development endpoint. AWS Glue provides an interface to SageMaker AI Jupyter notebooks. With AWS Glue, you create and manage SageMaker AI notebooks. You can also open SageMaker AI notebooks from the AWS Glue console.

In addition, you can use Apache Spark with SageMaker AI on AWS Glue development endpoints which support SageMaker AI (but not AWS Glue ETL jobs). SageMaker Spark is an open source Apache Spark library for SageMaker AI. For more information, see Using Apache Spark with Amazon SageMaker.

⚠ **Important**

Managing SageMaker AI notebooks with AWS Glue development endpoints is available in the following AWS Regions:

Region	**Code**
US East (Ohio)	`us-east-2`
US East (N. Virginia)	`us-east-1`
US West (N. California)	`us-west-1`
US West (Oregon)	`us-west-2`
Asia Pacific (Tokyo)	`ap-northeast-1`
Asia Pacific (Seoul)	`ap-northeast-2`
Asia Pacific (Mumbai)	`ap-south-1`
Asia Pacific (Singapore)	`ap-southeast-1`
Asia Pacific (Sydney)	`ap-southeast-2`
Canada (Central)	`ca-central-1`
Europe (Frankfurt)	`eu-central-1`
Europe (Ireland)	`eu-west-1`
Europe (London)	`eu-west-2`

Building visual ETL jobs

Build visual ETL jobs with AWS Glue Studio

AWS Glue Studio provides a visual interface for creating, running, and monitoring Extract/Transform/Load (ETL) jobs in AWS Glue. A job in AWS Glue consists of the business logic that performs extract, transform, and load (ETL) work. With AWS Glue Studio, you can visually compose data transformation workflows and seamlessly run them on AWS Glue's Apache Spark-based serverless ETL engine. You can create jobs that move and transform data between various data stores and streams using a drag-and-drop interface without having to learn Spark or write code.

An AWS Glue job encapsulates a script that connects to your source data, processes it, and then writes it out to your data target. Typically, a job runs extract, transform, and load (ETL) scripts. Jobs can run scripts designed for Apache Spark and Ray runtime environments. Jobs can also run general-purpose Python scripts (Python shell jobs.) AWS Glue triggers can start jobs based on a schedule or event, or on demand. You can monitor job runs to understand runtime metrics such as completion status, duration, and start time.

You can use scripts that AWS Glue generates or you can provide your own. With a source schema and target location or schema, the AWS Glue Studio code generator can automatically create an Apache Spark API (PySpark) script. You can use this script as a starting point and edit it to meet your goals.

AWS Glue can write output files in several data formats. Each job type may support different output formats. For some data formats, common compression formats can be written.

Managing AWS Glue Jobs in the AWS Console

To view existing jobs, sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/. Then choose the **Jobs** tab in AWS Glue. The **Jobs** list displays the location of the script that is associated with each job, when the job was last modified, and the current job bookmark option.

You can create jobs in the **ETL** section of the AWS Glue console. While creating a new job, or after you have saved your job, you can use can AWS Glue Studio to modify your ETL jobs. You can do this by editing the nodes in the visual editor or by editing the job script in developer mode. You can also add and remove nodes in the visual editor to create more complicated ETL jobs.

Next steps for creating a job in AWS Glue Studio

You use the visual job editor to configure nodes for your job. Each node represents an action, such as reading data from the source location or applying a transform to the data. Each node you add to your job has properties that provide information about either the data location or the transform.

The next steps for creating and managing your jobs are:

- Starting visual ETL jobs in AWS Glue Studio
- View the job script
- Modify the job properties
- Save the job
- Start a job run
- View information for recent job runs
- Accessing the job monitoring dashboard

Build visual ETL flows with Amazon SageMaker

With an Amazon SageMaker Unified Studio workflow, you can set up and run a series of tasks in Amazon SageMaker Unified Studio. Amazon SageMaker Unified Studio workflows use Apache Airflow to model data processing procedures and orchestrate your Amazon SageMaker Unified Studio code artifacts. For more information, see Using workflows in Amazon SageMaker Unified Studio .

Starting visual ETL jobs in AWS Glue Studio

You can use the simple visual interface in AWS Glue Studio to create your ETL jobs. You use the **Jobs** page to create new jobs. You can also use a script editor or notebook to work directly with code in the AWS Glue Studio ETL job script.

On the **Jobs** page, you can see all the jobs that you have created either with AWS Glue Studio or AWS Glue. You can view, manage, and run your jobs on this page.

Also see the blog tutorial
on another example of how to create ETL jobs with AWS Glue Studio.

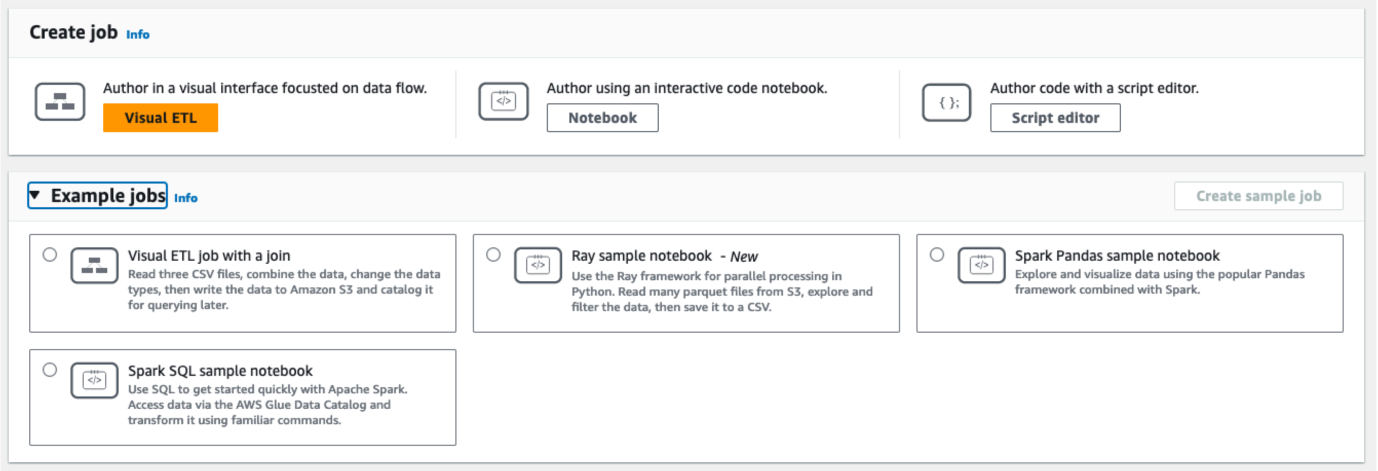
Starting jobs in AWS Glue Studio

AWS Glue allows you to create a job through a visual interface, an interactive code notebook, or with a script editor. You can start a job by clicking on any of the options or create a new job based on a sample job.

Sample jobs create a job with the tool of your choice. For example, sample jobs allow you to create a visual ETL job that joins CSV files into a catatlog table, create a job in an interactive code notebook with AWS Glue for Ray or AWS Glue for Spark when working with pandas, or create a job in an interactive code notebook with SparkSQL.

Creating a job in AWS Glue Studio from scratch

1. Sign in to the AWS Management Console and open the AWS Glue Studio console at https://console.aws.amazon.com/gluestudio/.
2. Choose **ETL jobs** from the navigation pane.
3. In the **Create job** section, select a configuration option for your job.



Author in a visual interface focusted on data flow.

Author using an interactive code notebook.

Author code with a script editor.

○ Visual ETL job with a join
Read three CSV files, combine the data, change the data types, then write the data to Amazon S3 and catalog it for querying later.

○ Ray sample notebook - *New*
Use the Ray framework for parallel processing in Python. Read many parquet files from S3, explore and filter the data, then save it to a CSV.

○ Spark Pandas sample notebook
Explore and visualize data using the popular Pandas framework combined with Spark.

○ Spark SQL sample notebook
Use SQL to get started quickly with Apache Spark. Access data via the AWS Glue Data Catalog and transform it using familiar commands.

Options to create a job from scratch:

- **Visual ETL** – author in a visual interface focused on data flow
- **Author using an Interactive code notebook** – interactively author jobs in a notebook interface based on Jupyter Notebooks

  When you select this option, you must provide additional information before creating a notebook authoring session. For more information about how to specify this information, see [Getting started with notebooks in AWS Glue Studio](#).

- **Author code with a script editor** – For those familiar with programming and writing ETL scripts, choose this option to create a new Spark ETL job. Choose the engine (Python shell, Ray, Spark (Python), or Spark (Scala). Then, choose **Start fresh** or **Upload script**. uploading an existing script from a local file. If you choose to use the script editor, you can't use the visual job editor to design or edit your job.

  A Spark job is run in an Apache Spark environment managed by AWS Glue. By default, new scripts are coded in Python. To write a new Scala script, see Creating and editing Scala scripts in AWS Glue Studio.

Creating a job in AWS Glue Studio from an example job

You can choose to create a job from an example job. In the **Example jobs** section, choose a sample job, then choose **Create sample job**. Creating a sample job from one of the options provides a quick template you can work from.

1. Sign in to the AWS Management Console and open the AWS Glue Studio console at https://console.aws.amazon.com/gluestudio/.
2. Choose **ETL jobs** from the navigation pane.
3. Select an option create a job from a sample job:
   - **Visual ETL job to join multiple sources** – Read three CSV files, combine the data, change the data types, then write the data to Amazon S3 and catalog it for querying later.
   - **Spark notebook using Pandas** – Explore and visualize data using the popular Pandas framework combined with Spark.
   - **Spark notebook using SQL** – Use SQL to get started quickly with Apache Spark. Access data through the AWS Glue Data Catalog and transform it using familiar commands.
4. Choose **Create sample job**.

Job editor features

The job editor provides the following features for creating and editing jobs.

- A visual diagram of your job, with a node for each job task: Data source nodes for reading the data; transform nodes for modifying the data; data target nodes for writing the data.

You can view and configure the properties of each node in the job diagram. You can also view the schema and sample data for each node in the job diagram. These features help you to verify that your job is modifying and transforming the data in the right way, without having to run the job.

- A Script viewing and editing tab, where you can modify the code generated for your job.
- A Job details tab, where you can configure a variety of settings to customize the environment in which your AWS Glue ETL job runs.
- A Runs tab, where you can view the current and previous runs of the job, view the status of the job run, and access the logs for the job run.
- A Data quality tab, where you can apply data quality rules to your job.
- A Schedules tab, where you can configure the start time for you job, or set up a recurring job runs.
- A Version Control tab, where you can configure a Git service to use with your job.

Using schema previews in the visual job editor

While creating or editing your job, you can use the **Output schema** tab to view the schema for your data.

Before you can see the schema, the job editor needs permissions to access the data source. You can specify an IAM role on the Job details tab of the editor or on the **Output schema** tab for a node. If the IAM role has all the necessary permissions to access the data source, you can then view the schema on the **Output schema** tab for a node.

Using data previews in the visual job editor

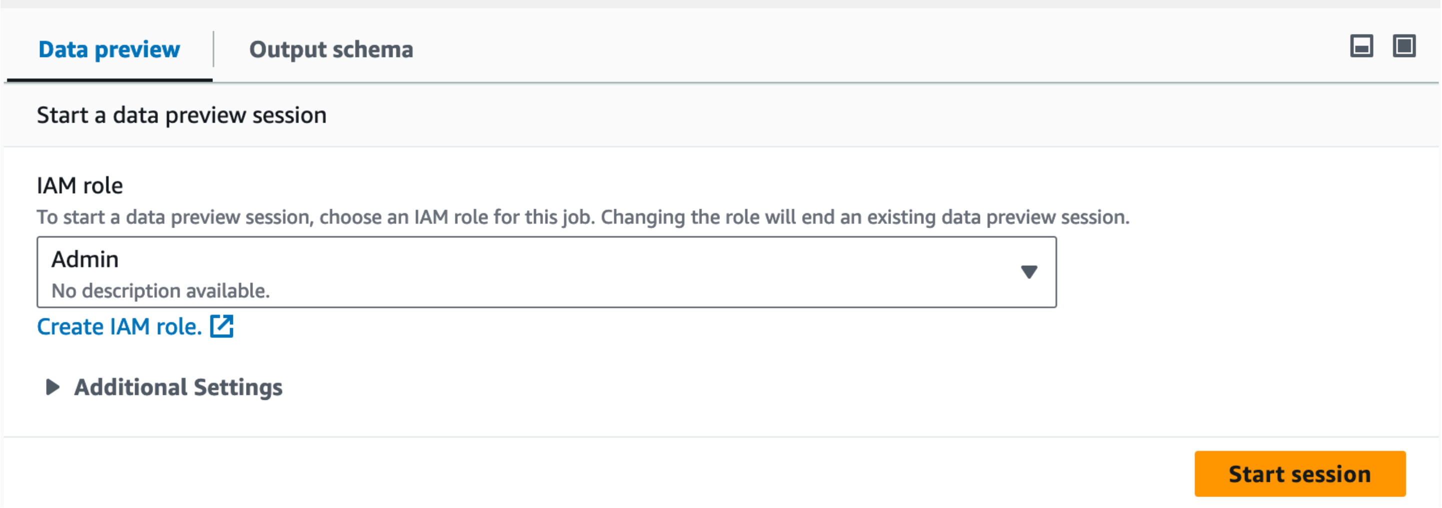
Data previews help you create and test your job using a sample of your data without having to repeatedly run the job. By using data preview, you can:

- Test an IAM role to make sure you have access to your data sources or data targets.
- Check that the transform is modifying the data in the intended way. For example, if you use a Filter transform, you can make sure that the filter is selecting the right subset of data.
- Check your data. If your dataset contains columns with values of multiple types, the data preview shows a list of tuples for these columns. Each tuple contains the data type and its value.

Note

If you use a data preview session and a custom SQL or custom code node, the data preview session will execute the SQL or code block as-is for the entire dataset.

While creating or editing your job, you can use the **Data preview** tab beneath the job canvas to view a sample of your data. A new data preview session will start automatically when the role is already configured on the job or a default IAM role has been set up in the account. If a role has not been previously configured, you can start a session by selecting the role.



Start a data preview session

IAM role

To start a data preview session, choose an IAM role for this job. Changing the role will end an existing data preview session.

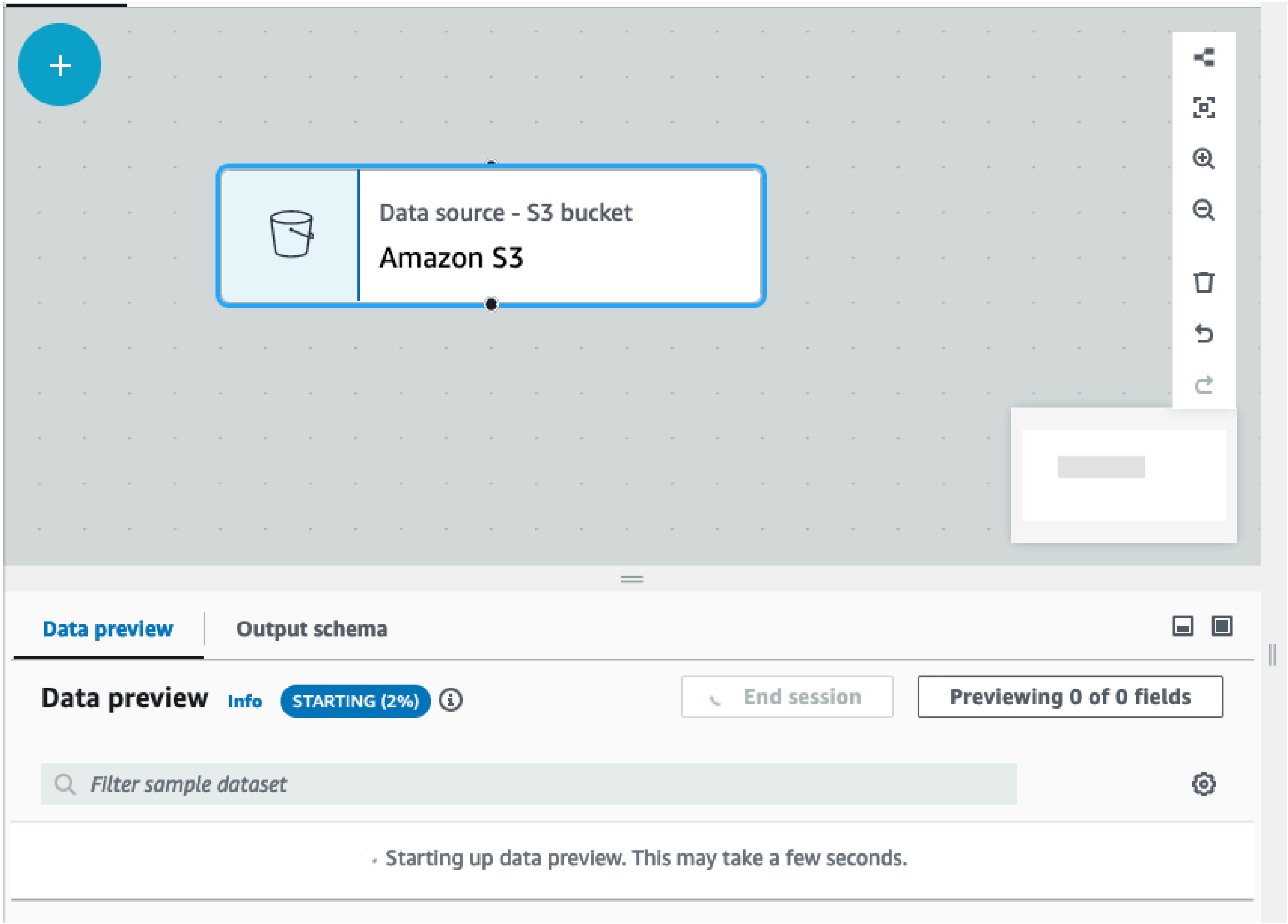
Admin
No description available.

Note

The role you choose for the data preview session will also be used for the job.

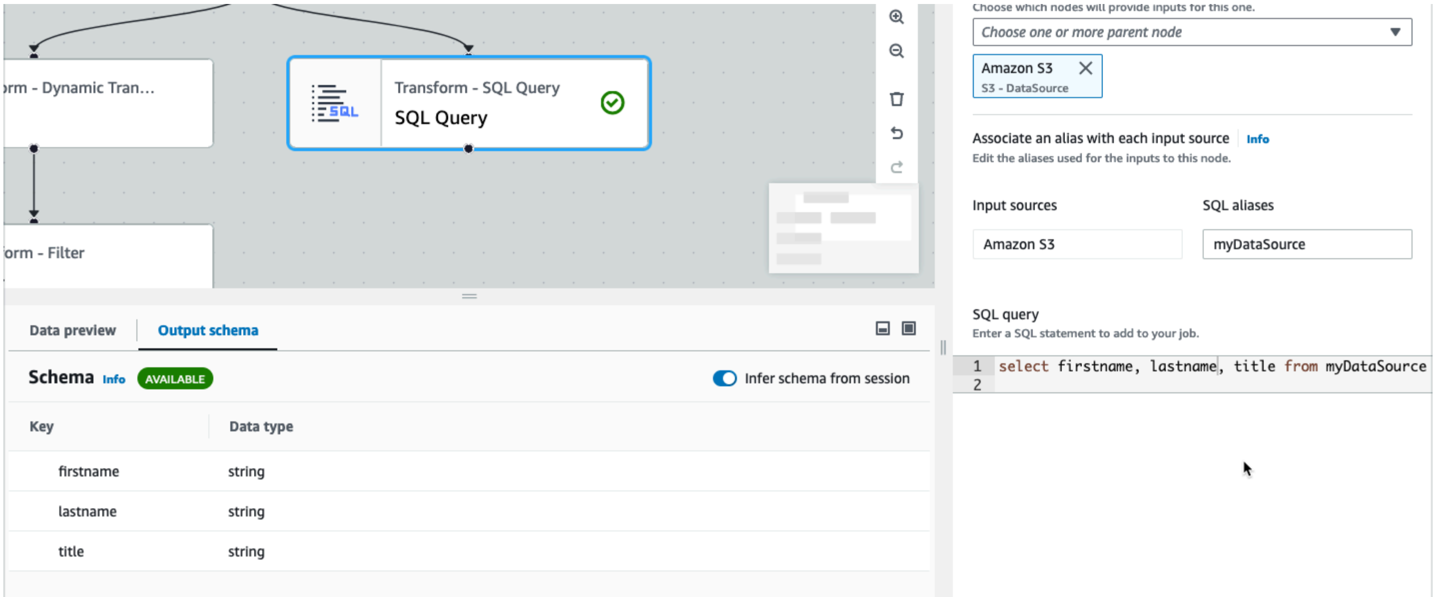
You can see the status and the progress of your session as well as the session details by clicking the info icon.

When the session is ready, AWS Glue Studio will load the data for the node you selected. You can view the **% complete** as it progresses.


Data source - S3 bucket
Amazon S3
Data preview
Output schema
Data preview
Info
STARTING (2%)
End session
Previewing 0 of 0 fields
Filter sample dataset

Starting up data preview. This may take a few seconds.

As you author your visual job, AWS Glue Studio will automatically update the schema for the selected node when you toggle **Infer schema from session** in the **Output schema** tab.



Choose which nodes will provide inputs for this one.

Choose one or more parent node

Amazon S3
S3 - DataSource

Associate an alias with each input source Info

Edit the aliases used for the inputs to this node.

Input sources	SQL aliases
Amazon S3	myDataSource

SQL query

Enter a SQL statement to add to your job.

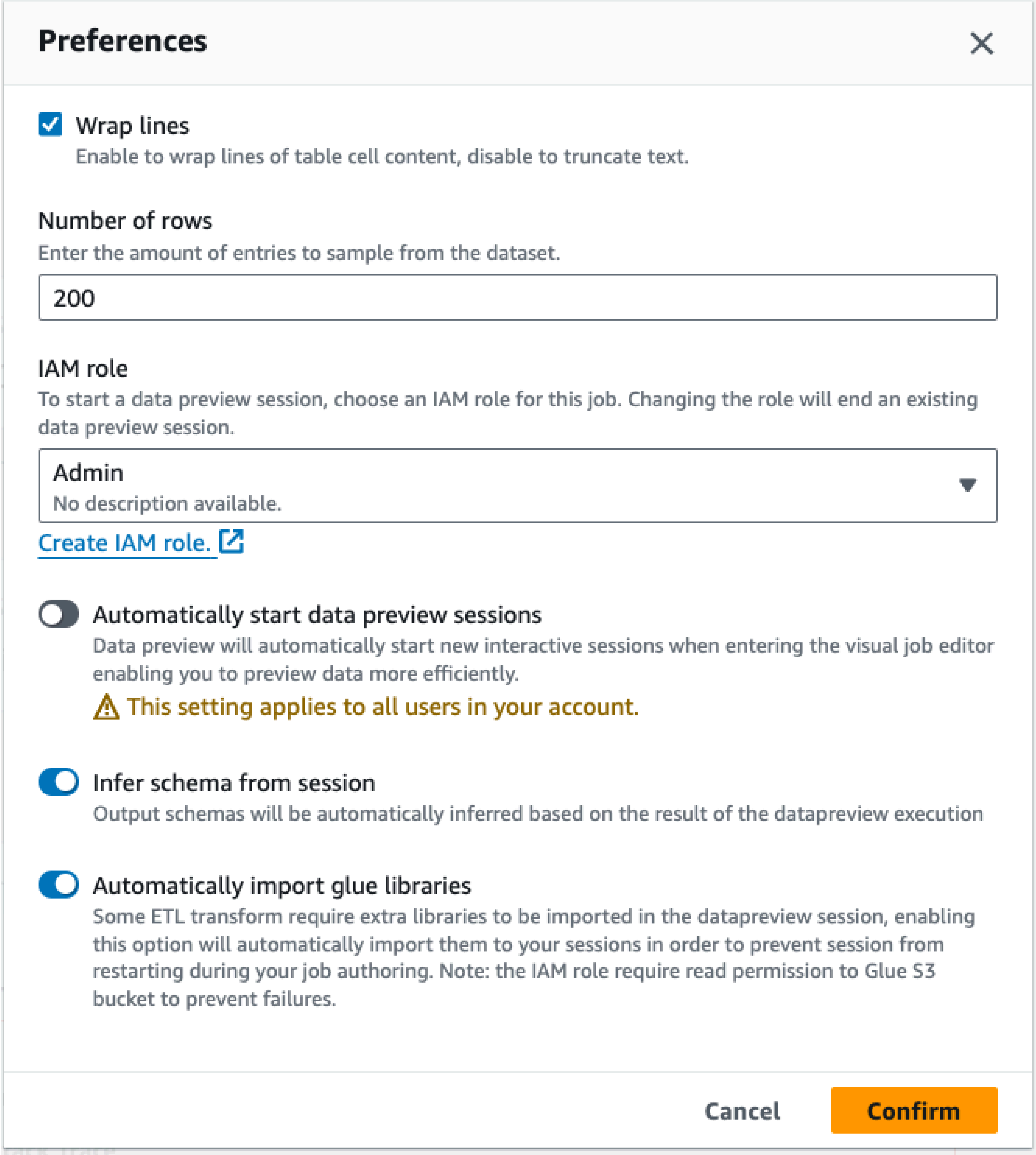
```
select firstname, lastname, title from myDataSource
```

Key	Data type
firstname	string
lastname	string
title	string

To configure your data preview preferences:

Choose the settings icon (a gear symbol) to configure your preferences for data previews. These settings apply to all nodes in the job diagram. You can:

- Choose to wrap the text from one line to the next. This option is enabled by default
- Change the number of rows (default to 200)
- Choose an IAM role or create an IAM role if needed
- Choose to automatically start a new session when you author a job. This provisions a new interactive session when authoring jobs. **This setting applies at the account level.** Once set, it will apply to all users in your account when editing any job.
- Choose to automatically infer schema. Output schemas will be automatically inferred for the selected node
- Choose to automatically import AWS Glue libraries. This is useful as it will prevent data preview from restarting new sessions when adding new transforms that require a session restart



Preferences

☑ Wrap lines
Enable to wrap lines of table cell content, disable to truncate text.

Number of rows
Enter the amount of entries to sample from the dataset.
200

IAM role
To start a data preview session, choose an IAM role for this job. Changing the role will end an existing data preview session.
Admin
No description available.
Create IAM role.

○ Automatically start data preview sessions
Data preview will automatically start new interactive sessions when entering the visual job editor enabling you to preview data more efficiently.
⚠ This setting applies to all users in your account.

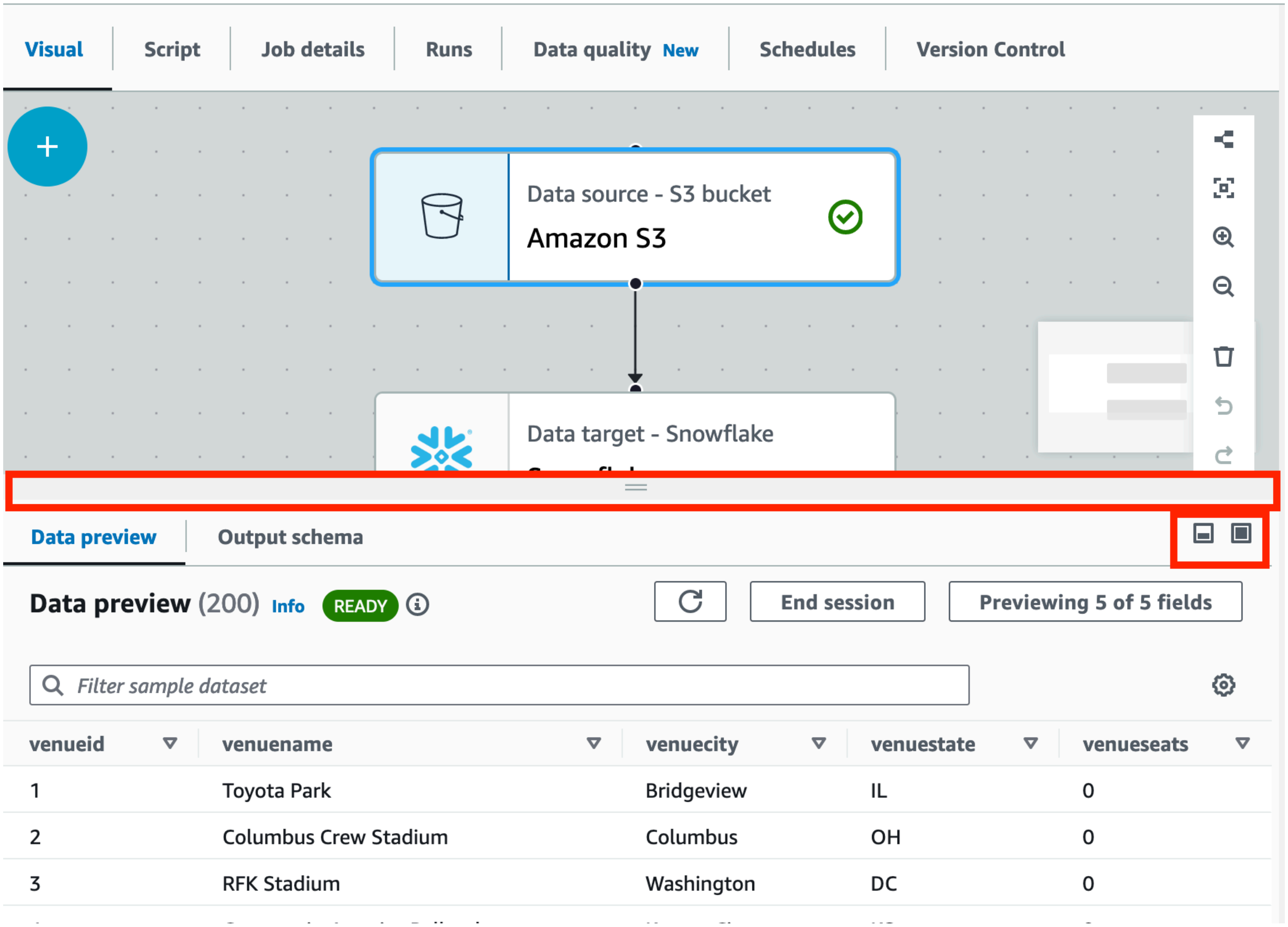
● Infer schema from session
Output schemas will be automatically inferred based on the result of the datapreview execution

● Automatically import glue libraries
Some ETL transform require extra libraries to be imported in the datapreview session, enabling this option will automatically import them to your sessions in order to prevent session from restarting during your job authoring. Note: the IAM role require read permission to Glue S3 bucket to prevent failures.

Additional features include the ability to:

- Choose the **Previewing x of y fields** button to select which columns (fields) to preview. When you preview you data using the default settings, the job editor shows the first 5 columns of your dataset. You can change this to show all or none (not recommended).

- Scroll through the data preview window both horizontally and vertically.
- Use the maximize button to expand the Data preview tab to over-lay the job graph to better view the data and data structures. Similarly, use the minimize button to minimize the Data preview tab. You can also grab the handle pane and drag up to expand the **Data preview** tab.


Visual
Script
Job details
Runs
Data quality New
Schedules
Version Control
Data source - S3 bucket
Amazon S3
Data target - Snowflake
Data preview
Output schema
Data preview (200) Info READY
End session
Previewing 5 of 5 fields
Filter sample dataset

venueid	venuename	venuecity	venuestate	venueseats
1	Toyota Park	Bridgeview	IL	0
2	Columbus Crew Stadium	Columbus	OH	0
3	RFK Stadium	Washington	DC	0

- Use **End session** to stop the data preview. When you stop the session, you can choose a new IAM role, and set additional settings (such as turn on or off settings to automatically start a new session, infer schema, or import AWS Glue libraries, and start the session again.

Restrictions when using data previews

When using data previews, you might encounter the following restrictions or limitations.

- The first time you choose the Data preview tab you must choose IAM role. This role must have the necessary permissions to access the data and other resources needed to create the data previews.

- After you provide an IAM role, it takes a while before the data is available for viewing. For datasets with less than 1 GB of data, it can take up to one minute. If you have a large dataset, you should use partitions to improve the loading time. Loading data directly from Amazon S3 has the best performance.
- If you have a very large dataset, and it takes more than 15 minutes to query the data for the data preview, the request will time out. Data previews have a 30 minute IDLE timeout. To alleviate this, reduce the dataset size to use data previews.
- By default, you see the first 50 columns in the Data preview tab. If the columns have no data values, you will get a message that there is no data to display. You can increase the number of rows sampled, or selected different columns to see data values.
- Data previews are currently not supported for streaming data sources, or for data sources that use custom connectors.
- Errors on one node effect the entire job. If any one node has an error with data previews, the error will show up on all nodes until you correct it.
- If you change a data source for the job, then the child nodes of that data source might need to be updated to match the new schema. For example, if you have an ApplyMapping node that modifies a column, and the column does not exist in the replacement data source, you will need to update the ApplyMapping transform node.
- If you view the Data preview tab for a SQL query transform node, and the SQL query uses an incorrect field name, the Data preview tab shows an error.

Script code generation

When you use the visual editor to create a job, the ETL code is automatically generated for you. AWS Glue Studio creates a functional and complete job script, and saves it in an Amazon S3 location.

There are two forms of code generated by AWS Glue Studio: the original, or Classic version, and a newer, streamlined version. By default, the new code generator is used to create the job script. You can generate a job script using classic code generator on the **Script** tab by choosing the **Generate classic script** toggle button.

Some of the differences in the new version of the generated code include:

- Large comment blocks are no longer added to the script

- Output structures in the code use the node name that you specify in the visual editor. In the class script, the output structures are simply named `DataSource0`, `DataSource1`, `Transform0`, `Transform1`, `DataSink0`, `DataSink1`, and so on.
- Long commands are split across multiple lines to remove the need to scroll across the page to see the entire command.

New features in AWS Glue Studio require the new version of code generation, and will not work with the classic code script. You are prompted to update these jobs when you attempt to run them.

Transform data with AWS Glue managed transforms

AWS Glue Studio provides two types of transforms:

- **AWS Glue-native transforms** - available to all users and are managed by AWS Glue.
- **Custom visual transforms** - allows you to upload your own transforms to use in AWS Glue Studio

AWS Glue managed data transform nodes

AWS Glue Studio provides a set of built-in transforms that you can use to process your data. Your data passes from one node in the job diagram to another in a data structure called a `DynamicFrame`, which is an extension to an Apache Spark SQL `DataFrame`.

In the pre-populated diagram for a job, between the data source and data target nodes is the **Change Schema** transform node. You can configure this transform node to modify your data, or you can use additional transforms.

The following built-in transforms are available with AWS Glue Studio:

- **ChangeSchema**: Map data property keys in the data source to data property keys in the data target. You can rename keys, modify the data types for keys, and choose which keys to drop from the dataset.
- **SelectFields**: Choose the data property keys that you want to keep.
- **DropFields**: Choose the data property keys that you want to drop.
- **RenameField**: Rename a single data property key.
- **Spigot**: Write samples of the data to an Amazon S3 bucket.

- **Join**: Join two datasets into one dataset using a comparison phrase on the specified data property keys. You can use inner, outer, left, right, left semi, and left anti joins.
- **Union**: Combine rows from more than one data source that have the same schema.
- **SplitFields**: Split data property keys into two `DynamicFrames`. Output is a collection of `DynamicFrames`: one with selected data property keys, and one with the remaining data property keys.
- **SelectFromCollection**: Choose one `DynamicFrame` from a collection of `DynamicFrames`. The output is the selected `DynamicFrame`.
- **FillMissingValues**: Locate records in the dataset that have missing values and add a new field with a suggested value that is determined by imputation
- **Filter**: Split a dataset into two, based on a filter condition.
- **Drop Null Fields**: Removes columns from the dataset if all values in the column are ‘null’.
- **Drop Duplicates**: Removes rows from your data source by choosing to match entire rows or specify keys.
- **SQL**: Enter SparkSQL code into a text entry field to use a SQL query to transform the data. The output is a single `DynamicFrame`.
- **Aggregate**: Performs a calculation (such as average, sum, min, max) on selected fields and rows, and creates a new field with the newly calculated value(s).
- **Flatten**: Extract fields inside structs into top level fields.
- **UUID**: Add a column with a Universally Unique Identifier for each row.
- **Identifier**: Add a column with a numeric identifier for each row.
- **To timestamp**: Convert a column to timestamp type.
- **Format timestamp**: Convert a timestamp column to a formatted string.
- **Conditional Router transform**: Apply multiple conditions to incoming data. Each row of the incoming data is evaluated by a group filter condition and processed into its corresponding group.
- **Concatenate Columns transform**: Build a new string column using the values of other columns with an optional spacer.
- **Split String transform**: Break up a string into an array of tokens using a regular expression to define how the split is done.
- **Array To Columns transform**: Extract some or all the elements of a column of type array into new columns.

- **Add Current Timestamp transform**: Mark the rows with the time on which the data was processed. This is useful for auditing purposes or to track latency in the data pipeline.
- **Pivot Rows to Columns transform**: Aggregate a numeric column by rotating unique values on selected columns which become new columns. If multiple columns are selected, the values are concatenated to name the new columns.
- **Unpivot Columns To Rows transform**: Convert columns into values of new columns generating a row for each unique value.
- **Autobalance Processing transform**: Redistribute the data better among the workers. This is useful where the data is unbalanced or as it comes from the source doesn't allow enough parallel processing on it.
- **Derived Column transform**: Define a new column based on a math formula or SQL expression in which you can use other columns in the data, as well as constants and literals.
- **Lookup transform**: Add columns from a defined catalog table when the keys match the defined lookup columns in the data.
- **Explode Array or Map Into Rows transform**: Extract values from a nested structure into individual rows that are easier to manipulate.
- **Record matching transform**: Invoke an existing Record Matching machine learning data classification transform.
- **Remove null rows transform**: Remove from the dataset rows that have all columns as null, or empty.
- **Parse JSON column transform**: Parse a string column containing JSON data and convert it to a struct or an array column, depending if the JSON is an object or an array, respectively.
- **Extract JSON path transform**: Extract new columns from a JSON string column.
- **Extract string fragments from a regular expression**: Extract string fragments using a regular expression and create new column out of it, or multiple columns if using regex groups.
- **Custom transform**: Enter code into a text entry field to use custom transforms. The output is a collection of `DynamicFrames`.

Using a data preparation recipe in AWS Glue Studio

The **Data preparation recipe** transform allows you to author a data preparation recipe from scratch using an interactive grid style authoring interface. It also allows you to import an existing AWS Glue DataBrew recipe and then edit it in AWS Glue Studio.

The **Data Preparation Recipe** node is available from the Resource panel. You can connect the **Data Preparation Recipe** node to another node in the visual workflow, whether it is a Data source node or another transformation node. After choosing a AWS Glue DataBrew recipe and version, the applied steps in the recipe are visible in the node properties tab.

Prerequisites

- If importing an AWS Glue DataBrew recipe, you have the required IAM permissions as described in Import a AWS Glue DataBrew recipe in AWS Glue Studio .
- A data preview session must be created.

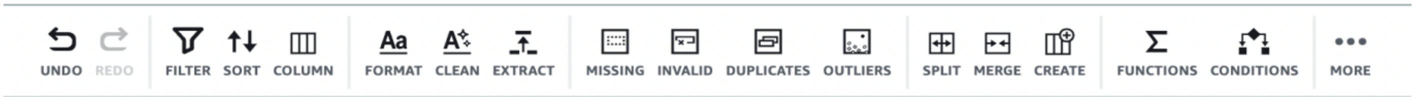
Limitations

- AWS Glue DataBrew recipes are only supported in commercial DataBrew regions.
- Not all AWS Glue DataBrew recipes are supported by AWS Glue. Some recipes will not be able to be run in AWS Glue Studio.
  - Recipes with `UNION` and `JOIN` transforms are not supported, however, AWS Glue Studio already has "Join" and "Union" transform nodes which can be used before or after a **Data Preparation Recipe** node.
- **Data Preparation Recipe** nodes are supported for jobs starting with AWS Glue version 4.0. This version will be auto-selected after a **Data Preparation Recipe** node is added to the job.
- **Data Preparation Recipe** nodes require Python. This is automatically set when the **Data Preparation Recipe** node is added to the job.
- Adding a new **Data Preparation Recipe** node to the visual graph will automatically restart your Data Preview session with the correct libraries to use the **Data Preparation Recipe** node.
- The following transforms are not supported for import or editing in a **Data Preparation Recipe** node: `GROUP_BY`, `PIVOT`, `UNPIVOT`, and `TRANSPOSE`.

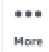
Additional features

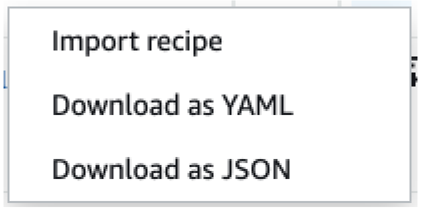
When you've selected the **Data Preparation Recipe** transform, you have the ability to take additional actions after choosing **Author recipe**.

- Add step – you can add additional steps to a recipe as needed by choosing the add step icon, or use the toolbar in the Preview pane by choosing an action.

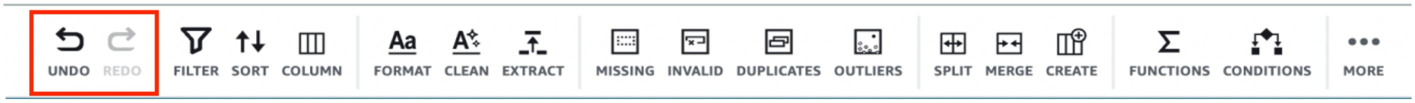

UNDO
REDO
FILTER
SORT
COLUMN
FORMAT
CLEAN
EXTRACT
MISSING
INVALID
DUPLICATES
OUTLIERS
SPLIT
MERGE
CREATE
FUNCTIONS
CONDITIONS
MORE

- Import recipe – choose **More** then **Import recipe** to use in your AWS Glue Studio job.




Import recipe
Download as YAML
Download as JSON

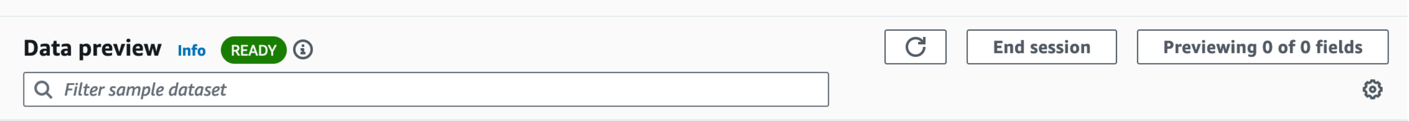
- Download as YAML – choose **More** then **Download as YAML** to download your recipe to save outside of AWS Glue Studio.
- Download as JSON – choose **More** then **Download as JSON** to download your recipe to save outside of AWS Glue Studio.
- Undo and redo recipe steps – You can undo and redo recipe steps in the Preview pane when working with data in the grid.


UNDO
REDO
FILTER
SORT
COLUMN
FORMAT
CLEAN
EXTRACT
MISSING
INVALID
DUPLICATES
OUTLIERS
SPLIT
MERGE
CREATE
FUNCTIONS
CONDITIONS
MORE

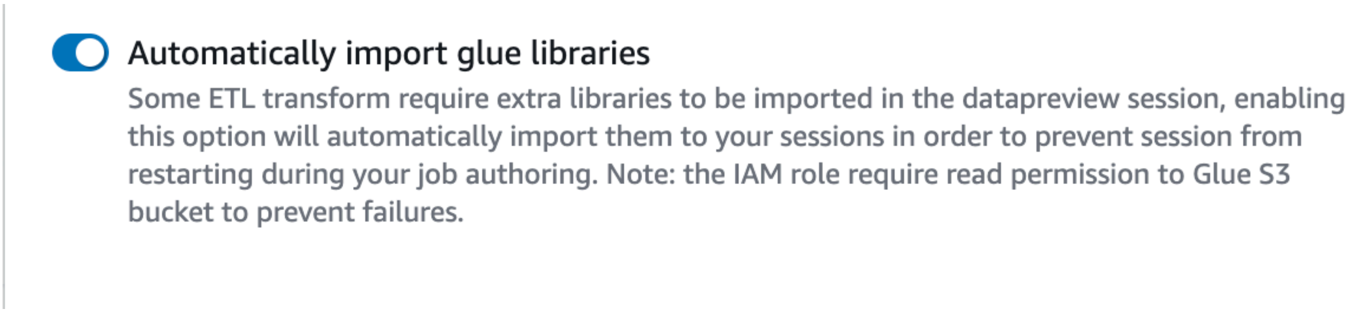
Author and run data preparation recipes in a visual ETL AWS Glue job

In this scenario, you can author data preparation recipes without having to first create them in DataBrew. Before you can start authoring recipes, you must:

- Have an active Data Preview session running. When the data preview session is READY, then **Author Recipe** will become active and you can begin authoring or editing your recipe.


Data preview
Info
READY
End session
Previewing 0 of 0 fields
Filter sample dataset

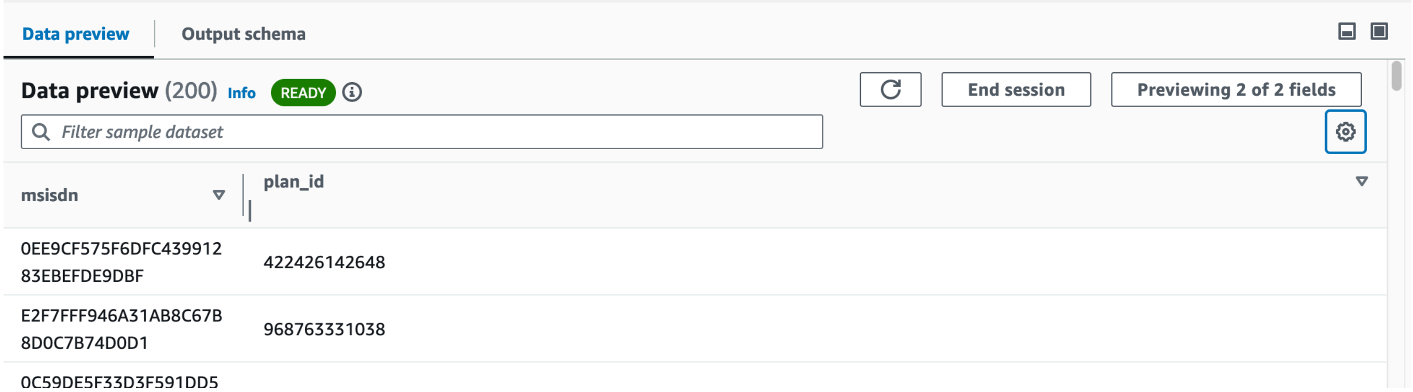
- Ensure that the toggle for **Automatically import glue libraries** is enabled.



Automatically import glue libraries

Some ETL transform require extra libraries to be imported in the datapreview session, enabling this option will automatically import them to your sessions in order to prevent session from restarting during your job authoring. Note: the IAM role require read permission to Glue S3 bucket to prevent failures.

You can do this by choosing the gear icon in the Data Preview pane.

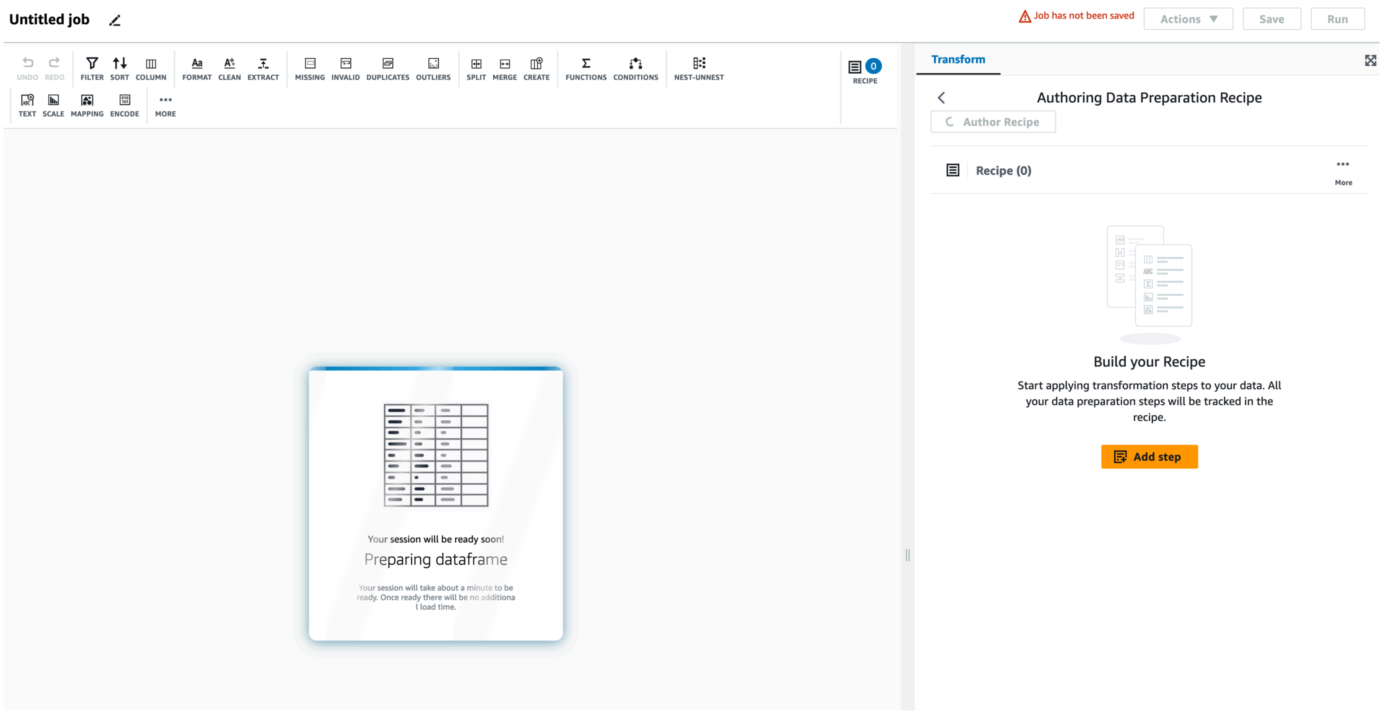


Data preview
Output schema
Data preview (200)
Info
READY
End session
Previewing 2 of 2 fields
Filter sample dataset

msisdn	plan_id
0EE9CF575F6DFC439912 83EBEFDE9DBF	422426142648
E2F7FFF946A31AB8C67B 8D0C7B74D0D1	968763331038
0C59DE5F33D3F591DD5	

To author a data preparation recipe in AWS Glue Studio:

1. Add the **Data Preparation Recipe** transform to your job canvas. Your transform should be connected to a data source node parent. When adding the **Data Preparation Recipe** node, the node will restart with the proper libraries and you will see the Data Frame being prepared.


Untitled job
Job has not been saved
Actions
Save
Run
UNDO
REDO
FILTER
SORT
COLUMN
FORMAT
CLEAN
EXTRACT
MISSING
INVALID
DUPLICATES
OUTLIERS
SPLIT
MERGE
CREATE
FUNCTIONS
CONDITIONS
NEST-UNNEST
TEXT
SCALE
MAPPING
ENCODE
MORE
0
RECIPE
Transform
Authoring Data Preparation Recipe
Author Recipe
Recipe (0)
More
Build your Recipe
Add step

Your session will be ready soon!

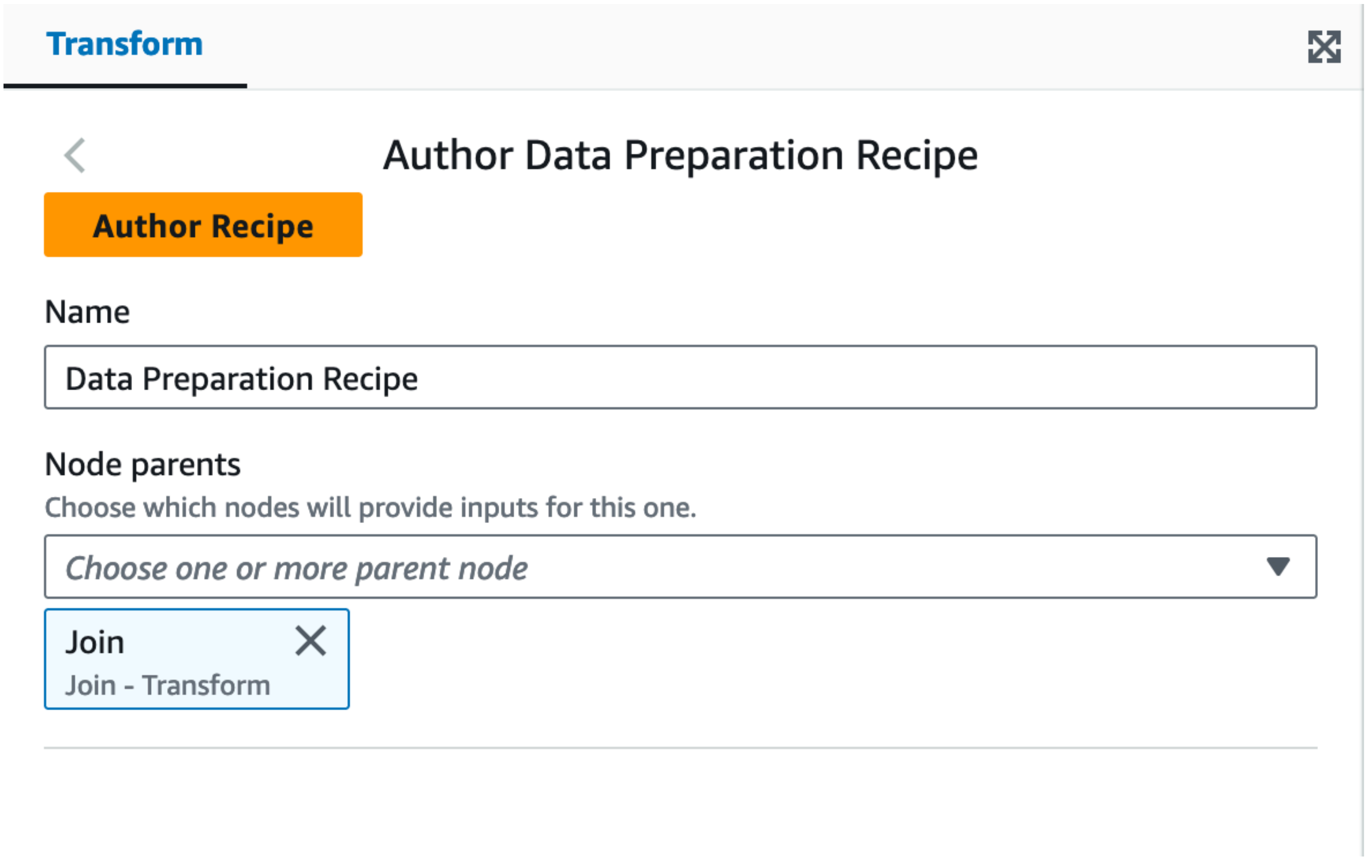
Preparing dataframe

Your session will take about a minute to be ready. Once ready there will be no additional load time.

Start applying transformation steps to your data. All your data preparation steps will be tracked in the recipe.

2. Once the Data Preview session is ready, the data with any previously applied steps will appear on the bottom of the screen.

3. Choose **Author Recipe**. This will allow you to start a new recipe in AWS Glue Studio.


Transform
Author Data Preparation Recipe
Author Recipe

Name

Data Preparation Recipe

Node parents

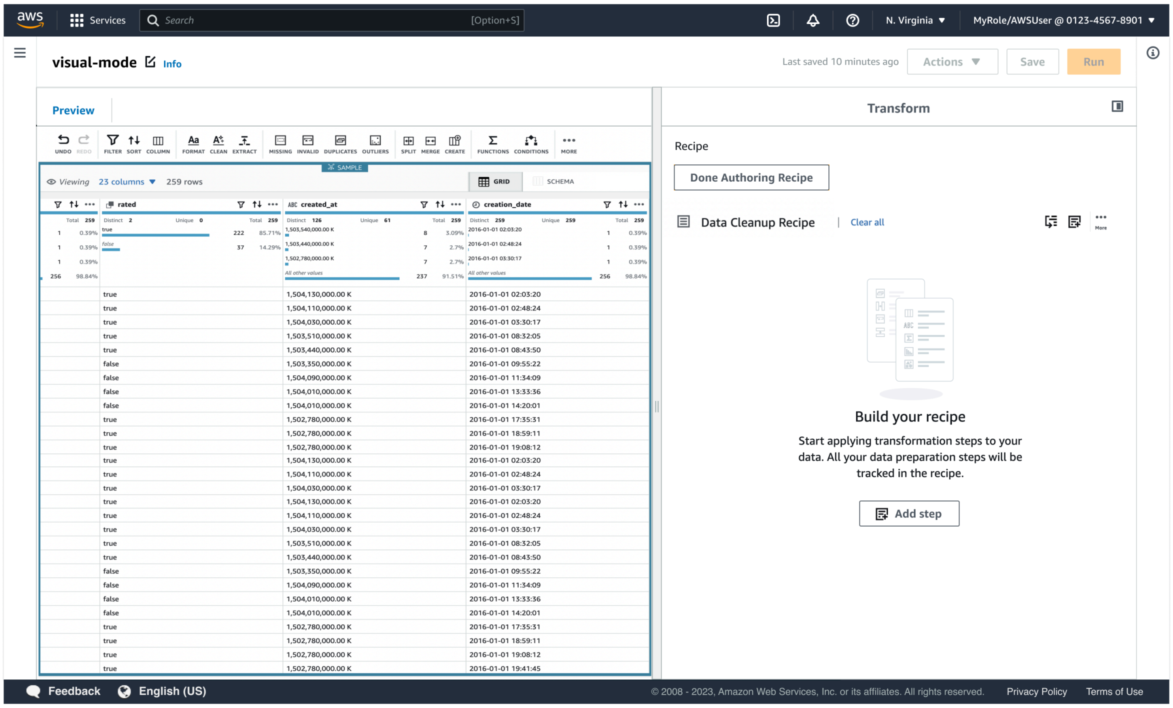
Choose which nodes will provide inputs for this one.

Choose one or more parent node

Join

Join - Transform

4. In the **Transform** panel to the right of the job canvas, enter a name for your data preparation recipe.
5. On the left-side, the canvas will be replaced with a grid view of your data. To the right, the **Transform** panel will change to show you your recipe steps. Choose **Add step** to add the first step in your recipe.


aws
Services
Search
[Option+S]
N. Virginia
MyRole/AWSUser @ 0123-4567-8901
Preview
UNDO
REDO
FILTER
SORT
COLUMN
FORMAT
CLEAN
EXTRACT
MISSING
INVALID
DUPLICATES
OUTLIERS
SPLIT
MERGE
CREATE
FUNCTIONS
CONDITIONS
MORE
SAMPLE
GRID
SCHEMA
Actions
Save
Run
Done Authoring Recipe
Clear all
More
Add step
Feedback
English (US)
© 2008 - 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

visual-mode Info

Last saved 10 minutes ago

Viewing 23 columns 259 rows

	rated	created_at	creation_date
Total 259	Distinct 2 Unique 0 Total 259	Distinct 126 Unique 61 Total 259	Distinct 259 Unique 259 Total 259
1 0.39%	true 222 85.71%	1,503,540,000.00 K 8 3.09%	2016-01-01 02:03:20 1 0.39%
1 0.39%	false 37 14.29%	1,503,440,000.00 K 7 2.7%	2016-01-01 02:48:24 1 0.39%
1 0.39%		1,502,780,000.00 K 7 2.7%	2016-01-01 03:30:17 1 0.39%
256 98.84%		All other values 237 91.51%	All other values 256 98.84%
	true	1,504,130,000.00 K	2016-01-01 02:03:20
	true	1,504,110,000.00 K	2016-01-01 02:48:24
	true	1,504,030,000.00 K	2016-01-01 03:30:17
	true	1,503,510,000.00 K	2016-01-01 08:32:05
	true	1,503,440,000.00 K	2016-01-01 08:43:50
	false	1,503,350,000.00 K	2016-01-01 09:55:22
	false	1,504,090,000.00 K	2016-01-01 11:34:09
	false	1,504,010,000.00 K	2016-01-01 13:33:36
	false	1,504,010,000.00 K	2016-01-01 14:20:01
	true	1,502,780,000.00 K	2016-01-01 17:35:31
	true	1,502,780,000.00 K	2016-01-01 18:59:11
	true	1,502,780,000.00 K	2016-01-01 19:08:12
	true	1,504,130,000.00 K	2016-01-01 02:03:20
	true	1,504,110,000.00 K	2016-01-01 02:48:24
	true	1,504,030,000.00 K	2016-01-01 03:30:17
	true	1,504,130,000.00 K	2016-01-01 02:03:20
	true	1,504,110,000.00 K	2016-01-01 02:48:24
	true	1,504,030,000.00 K	2016-01-01 03:30:17
	true	1,503,510,000.00 K	2016-01-01 08:32:05
	true	1,503,440,000.00 K	2016-01-01 08:43:50
	false	1,503,350,000.00 K	2016-01-01 09:55:22
	false	1,504,090,000.00 K	2016-01-01 11:34:09
	false	1,504,010,000.00 K	2016-01-01 13:33:36
	false	1,504,010,000.00 K	2016-01-01 14:20:01
	true	1,502,780,000.00 K	2016-01-01 17:35:31
	true	1,502,780,000.00 K	2016-01-01 18:59:11
	true	1,502,780,000.00 K	2016-01-01 19:08:12
	true	1,502,780,000.00 K	2016-01-01 19:41:45

Transform

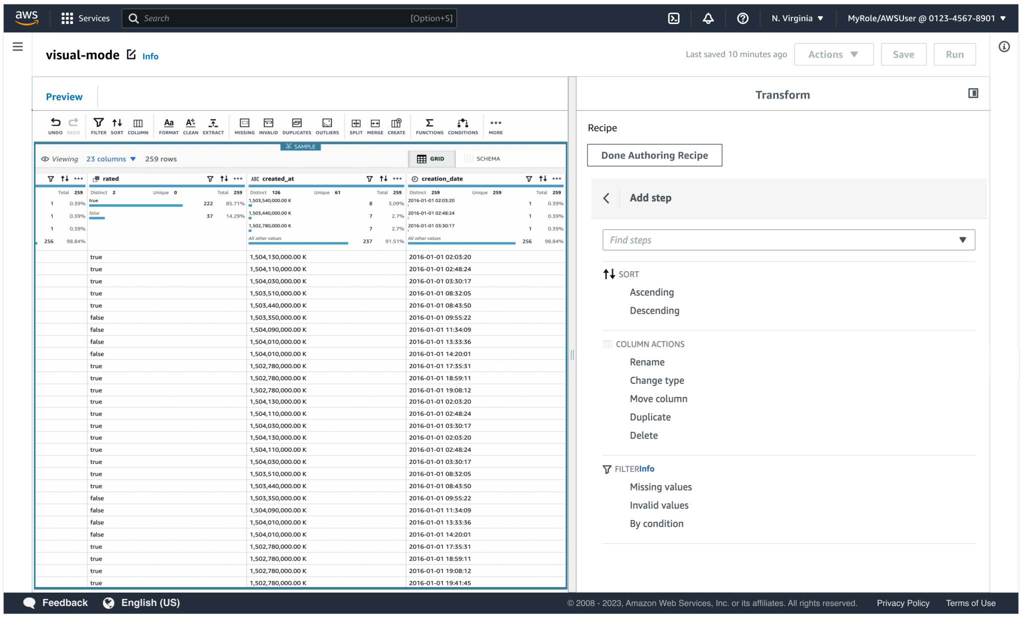
Recipe

Data Cleanup Recipe

Build your recipe

Start applying transformation steps to your data. All your data preparation steps will be tracked in the recipe.

6. In the **Transform** panel, choose to sort, take an action on the column, and filter values. For example, choose **Rename column**.

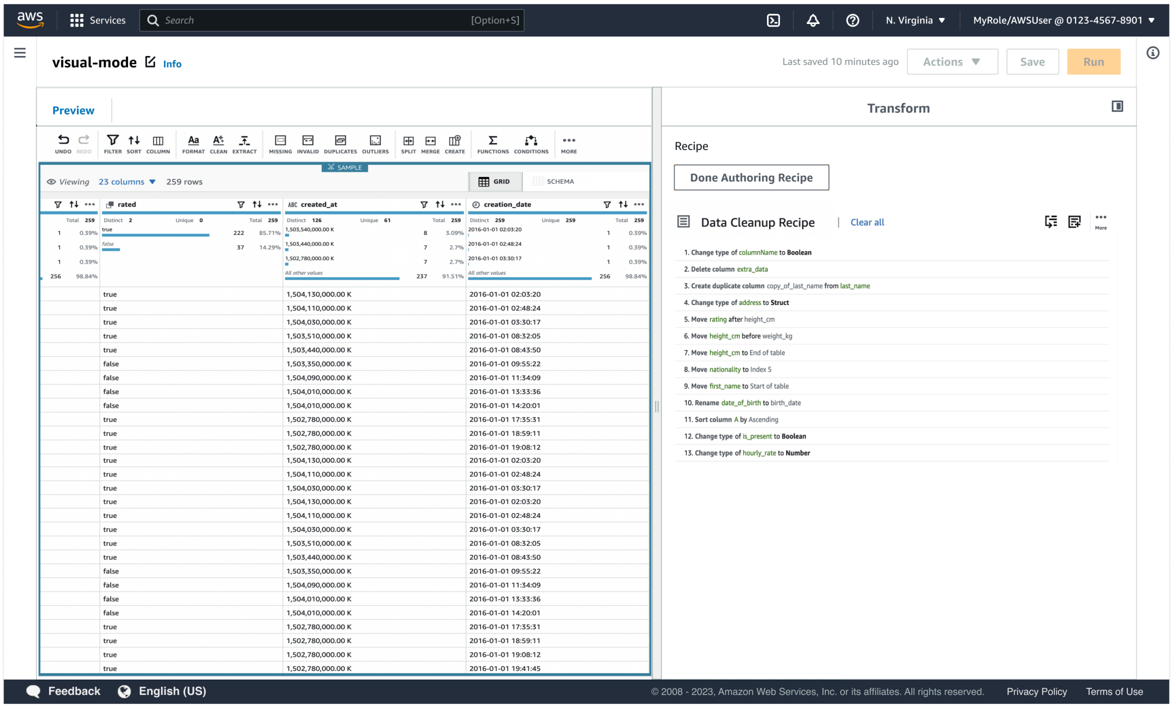

aws
Services
Search
[Option+S]
N. Virginia
MyRole/AWSUser @ 0123-4567-8901
visual-mode
Info
Last saved 10 minutes ago
Actions
Save
Run
Preview
UNDO
REDO
FILTER
SORT
COLUMN
FORMAT
CLEAN
EXTRACT
MISSING
INVALID
DUPLICATES
OUTLIERS
SPLIT
MERGE
CREATE
FUNCTIONS
CONDITIONS
MORE
SAMPLE
GRID
SCHEMA
Transform
Recipe
Done Authoring Recipe
Add step
Find steps
SORT
Ascending
Descending
COLUMN ACTIONS
Rename
Change type
Move column
Duplicate
Delete
FILTER Info
Missing values
Invalid values
By condition
Feedback
English (US)
© 2008 - 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

Viewing 23 columns 259 rows

	rated	created_at	creation_date
Total 259	Distinct 2 Unique 0 Total 259	Distinct 126 Unique 61 Total 259	Distinct 259 Unique 259 Total 259
1 0.39%	true 222 85.71%	1,503,540,000.00 K 8 3.09%	2016-01-01 02:03:20 1 0.39%
1 0.39%	false 37 14.29%	1,503,440,000.00 K 7 2.7%	2016-01-01 02:48:24 1 0.39%
1 0.39%		1,502,780,000.00 K 7 2.7%	2016-01-01 03:30:17 1 0.39%
256 98.84%		All other values 237 91.51%	All other values 256 98.84%
	true	1,504,130,000.00 K	2016-01-01 02:03:20
	true	1,504,110,000.00 K	2016-01-01 02:48:24
	true	1,504,030,000.00 K	2016-01-01 03:30:17
	true	1,503,510,000.00 K	2016-01-01 08:32:05
	true	1,503,440,000.00 K	2016-01-01 08:43:50
	false	1,503,350,000.00 K	2016-01-01 09:55:22
	false	1,504,090,000.00 K	2016-01-01 11:34:09
	false	1,504,010,000.00 K	2016-01-01 13:33:36
	false	1,504,010,000.00 K	2016-01-01 14:20:01
	true	1,502,780,000.00 K	2016-01-01 17:35:31
	true	1,502,780,000.00 K	2016-01-01 18:59:11
	true	1,502,780,000.00 K	2016-01-01 19:08:12
	true	1,504,130,000.00 K	2016-01-01 02:03:20
	true	1,504,110,000.00 K	2016-01-01 02:48:24
	true	1,504,030,000.00 K	2016-01-01 03:30:17
	true	1,504,130,000.00 K	2016-01-01 02:03:20
	true	1,504,110,000.00 K	2016-01-01 02:48:24
	true	1,504,030,000.00 K	2016-01-01 03:30:17
	true	1,503,510,000.00 K	2016-01-01 08:32:05
	true	1,503,440,000.00 K	2016-01-01 08:43:50
	false	1,503,350,000.00 K	2016-01-01 09:55:22
	false	1,504,090,000.00 K	2016-01-01 11:34:09
	false	1,504,010,000.00 K	2016-01-01 13:33:36
	false	1,504,010,000.00 K	2016-01-01 14:20:01
	true	1,502,780,000.00 K	2016-01-01 17:35:31
	true	1,502,780,000.00 K	2016-01-01 18:59:11
	true	1,502,780,000.00 K	2016-01-01 19:08:12
	true	1,502,780,000.00 K	2016-01-01 19:41:45

7. In the Transform panel on the right-side, options for renaming a column allow you to choose the source column to rename, and to enter the new column name. Once you have done so, choose **Apply**.

   You can preview each step, undo a step, and re-order steps and use any of the action icons, such as Filter, Sort, Split, Merge, etc. When you perform actions in the data grid, the steps are added to the recipe in the Transform panel.


aws
Services
Search
[Option+S]
N. Virginia
MyRole/AWSUser @ 0123-4567-8901
visual-mode
Info
Last saved 10 minutes ago
Actions
Save
Run
Preview
UNDO
REDO
FILTER
SORT
COLUMN
FORMAT
CLEAN
EXTRACT
MISSING
INVALID
DUPLICATES
OUTLIERS
SPLIT
MERGE
CREATE
FUNCTIONS
CONDITIONS
MORE
SAMPLE
GRID
SCHEMA
Transform
Clear all
More
Feedback
English (US)
© 2008 - 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

Viewing 23 columns 259 rows

	rated	created_at	creation_date
Total 259	Distinct 2 Unique 0 Total 259	Distinct 126 Unique 61 Total 259	Distinct 259 Unique 259 Total 259
1 0.39%	true 222 85.71%	1,503,540,000.00 K 8 3.09%	2016-01-01 02:03:20 1 0.39%
1 0.39%	false 37 14.29%	1,503,440,000.00 K 7 2.7%	2016-01-01 02:48:24 1 0.39%
1 0.39%		1,502,780,000.00 K 7 2.7%	2016-01-01 03:30:17 1 0.39%
256 98.84%		All other values 237 91.51%	All other values 256 98.84%
	true	1,504,130,000.00 K	2016-01-01 02:03:20
	true	1,504,110,000.00 K	2016-01-01 02:48:24
	true	1,504,030,000.00 K	2016-01-01 03:30:17
	true	1,503,510,000.00 K	2016-01-01 08:32:05
	true	1,503,440,000.00 K	2016-01-01 08:43:50
	false	1,503,350,000.00 K	2016-01-01 09:55:22
	false	1,504,090,000.00 K	2016-01-01 11:34:09
	false	1,504,010,000.00 K	2016-01-01 13:33:36
	false	1,504,010,000.00 K	2016-01-01 14:20:01
	true	1,502,780,000.00 K	2016-01-01 17:35:31
	true	1,502,780,000.00 K	2016-01-01 18:59:11
	true	1,502,780,000.00 K	2016-01-01 19:08:12
	true	1,504,130,000.00 K	2016-01-01 02:03:20
	true	1,504,110,000.00 K	2016-01-01 02:48:24
	true	1,504,030,000.00 K	2016-01-01 03:30:17
	true	1,504,130,000.00 K	2016-01-01 02:03:20
	true	1,504,110,000.00 K	2016-01-01 02:48:24
	true	1,504,030,000.00 K	2016-01-01 03:30:17
	true	1,503,510,000.00 K	2016-01-01 08:32:05
	true	1,503,440,000.00 K	2016-01-01 08:43:50
	false	1,503,350,000.00 K	2016-01-01 09:55:22
	false	1,504,090,000.00 K	2016-01-01 11:34:09
	false	1,504,010,000.00 K	2016-01-01 13:33:36
	false	1,504,010,000.00 K	2016-01-01 14:20:01
	true	1,502,780,000.00 K	2016-01-01 17:35:31
	true	1,502,780,000.00 K	2016-01-01 18:59:11
	true	1,502,780,000.00 K	2016-01-01 19:08:12
	true	1,502,780,000.00 K	2016-01-01 19:41:45

Recipe

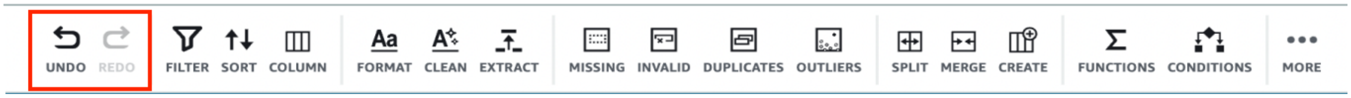
Done Authoring Recipe

Data Cleanup Recipe

1. Change type of columnName to Boolean
2. Delete column extra_data
3. Create duplicate column copy_of_last_name from last_name
4. Change type of address to Struct
5. Move rating after height_cm
6. Move height_cm before weight_kg
7. Move height_cm to End of table
8. Move nationality to Index 5
9. Move first_name to Start of table
10. Rename date_of_birth to birth_date
11. Sort column A by Ascending
12. Change type of is_present to Boolean
13. Change type of hourly_rate to Number

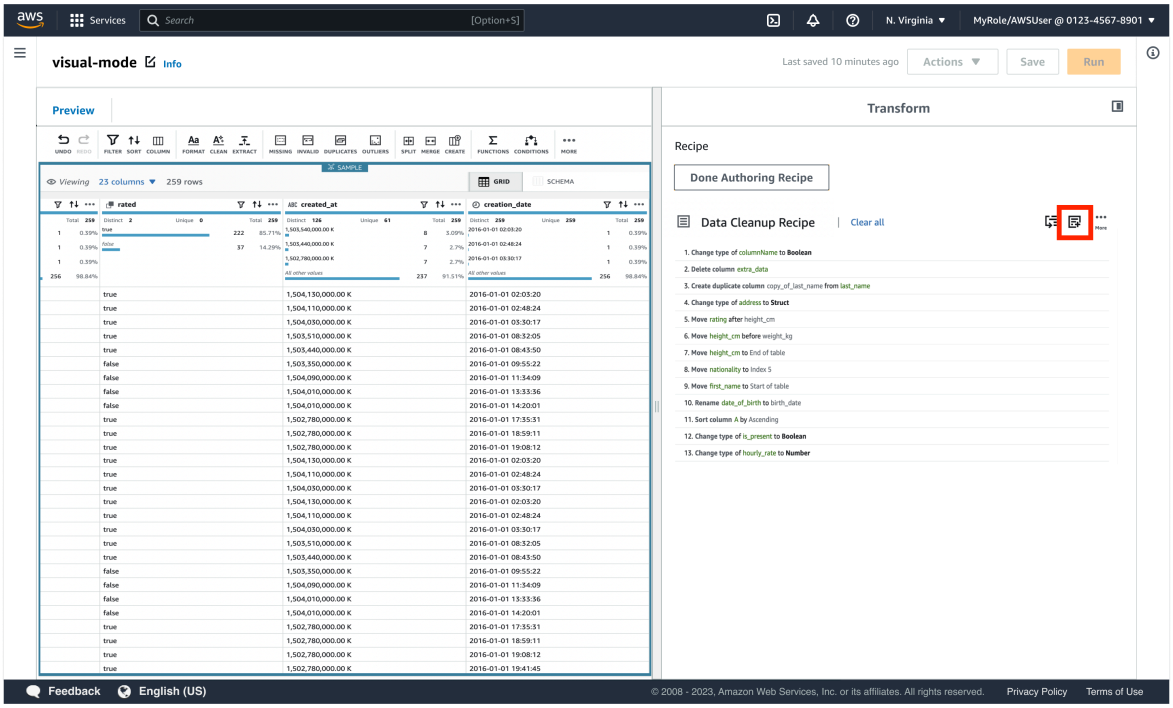
If you need to make a change, you can do this in the Preview pane by previewing the result of each step, undoing a step, and re-ordering steps. For example:

- Undo/redo step – undo a step by choosing the **undo** icon. You can repeat a step by choosing the **redo** icon.


UNDO
REDO
FILTER
SORT
COLUMN
FORMAT
CLEAN
EXTRACT
MISSING
INVALID
DUPLICATES
OUTLIERS
SPLIT
MERGE
CREATE
FUNCTIONS
CONDITIONS
MORE

- Reorder step – when you reorder a step, AWS Glue Studio will validate each step and let you know if the step is invalid.

8. Once you've applied a step, the Transform panel will show you all the steps in your recipe. You can clear all the steps to start over, add more steps by choosing the add icon, or choose **Done Authoring Recipe**.


Services
Search
[Option+S]
N. Virginia
MyRole/AWSUser @ 0123-4567-8901
visual-mode
Info
Last saved 10 minutes ago
Actions
Save
Run
Preview
UNDO
REDO
FILTER
SORT
COLUMN
FORMAT
CLEAN
EXTRACT
MISSING
INVALID
DUPLICATES
OUTLIERS
SPLIT
MERGE
CREATE
FUNCTIONS
CONDITIONS
MORE
SAMPLE
Viewing 23 columns 259 rows
GRID
SCHEMA
Transform
Recipe
Done Authoring Recipe
Data Cleanup Recipe
Clear all
More
Feedback
English (US)
© 2008 - 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

	rated	created_at	creation_date
	true	1,504,130,000.00 K	2016-01-01 02:03:20
	true	1,504,110,000.00 K	2016-01-01 02:48:24
	true	1,504,030,000.00 K	2016-01-01 03:30:17
	true	1,503,510,000.00 K	2016-01-01 08:32:05
	true	1,503,440,000.00 K	2016-01-01 08:43:50
	false	1,503,350,000.00 K	2016-01-01 09:55:22
	false	1,504,090,000.00 K	2016-01-01 11:34:09
	false	1,504,010,000.00 K	2016-01-01 13:33:36
	false	1,504,010,000.00 K	2016-01-01 14:20:01
	true	1,502,780,000.00 K	2016-01-01 17:35:31
	true	1,502,780,000.00 K	2016-01-01 18:59:11
	true	1,502,780,000.00 K	2016-01-01 19:08:12
	true	1,504,130,000.00 K	2016-01-01 02:03:20
	true	1,504,110,000.00 K	2016-01-01 02:48:24
	true	1,504,030,000.00 K	2016-01-01 03:30:17
	true	1,504,130,000.00 K	2016-01-01 02:03:20
	true	1,504,110,000.00 K	2016-01-01 02:48:24
	true	1,504,030,000.00 K	2016-01-01 03:30:17
	true	1,503,510,000.00 K	2016-01-01 08:32:05
	true	1,503,440,000.00 K	2016-01-01 08:43:50
	false	1,503,350,000.00 K	2016-01-01 09:55:22
	false	1,504,090,000.00 K	2016-01-01 11:34:09
	false	1,504,010,000.00 K	2016-01-01 13:33:36
	false	1,504,010,000.00 K	2016-01-01 14:20:01
	true	1,502,780,000.00 K	2016-01-01 17:35:31
	true	1,502,780,000.00 K	2016-01-01 18:59:11
	true	1,502,780,000.00 K	2016-01-01 19:08:12
	true	1,502,780,000.00 K	2016-01-01 19:41:45

1. Change type of columnName to Boolean
2. Delete column extra_data
3. Create duplicate column copy_of_last_name from last_name
4. Change type of address to Struct
5. Move rating after height_cm
6. Move height_cm before weight_kg
7. Move height_cm to End of table
8. Move nationality to Index 5
9. Move first_name to Start of table
10. Rename date_of_birth to birth_date
11. Sort column A by Ascending
12. Change type of is_present to Boolean
13. Change type of hourly_rate to Number

9. Choose **Save** at the top right side of your screen. Your recipe steps will not be saved until you save your job.

Import a AWS Glue DataBrew recipe in AWS Glue Studio

In AWS Glue DataBrew, a recipe is a set of data transformation steps. AWS Glue DataBrew recipes prescribes how to transform data that have already been read and doesn't describe where and how to read data, as well as how and where to write data. This is configured in Source and Target nodes in AWS Glue Studio. For more information on recipes, see Creating and using AWS Glue DataBrew recipes .

To use AWS Glue DataBrew recipes in AWS Glue Studio, begin with creating recipes in AWS Glue DataBrew. If you have recipes you want to use, you can skip this step.

IAM permissions for AWS Glue DataBrew

This topic provides information to help you understand the actions and resources that you an IAM administrator can use in an AWS Identity and Access Management (IAM) policy for the Data Preparation Recipe transform.

For additional information about security in AWS Glue, see Access Management.

> **Note**
>
> The following table lists the permissions that a user needs if importing an existing AWS Glue DataBrew recipe.

Data Preparation Recipe transform actions

Action	Description
`databrew:ListRecipes`	Grants permission to retrieve AWS Glue DataBrew recipes.
`databrew:ListRecipeVersions`	Grants permission to retrieve AWS Glue DataBrew recipe versions.
`databrew:DescribeRecipe`	Grants permission to retrieve AWS Glue DataBrew recipe description.

The role you're using for accessing this functionality should have a policy that allows several AWS Glue DataBrew actions. You can achieve this by either using the `AWSGlueConsoleFullAccess` policy that includes the necessary actions or add the following inline policy to your role:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
          "databrew:ListRecipes",
          "databrew:ListRecipeVersions",
          "databrew:DescribeRecipe"
        ],
        "Resource": [
          "*"
        ]
      }
    ]
  }
```

To use the Data Preparation Recipe transform, you must add the `IAM:PassRole` action to the permissions policy.

Additional required permissions

Action	Description
`iam:PassRole`	Grants permission for IAM to allow the user to pass the approved roles.

Without these permissions the following error occurs:

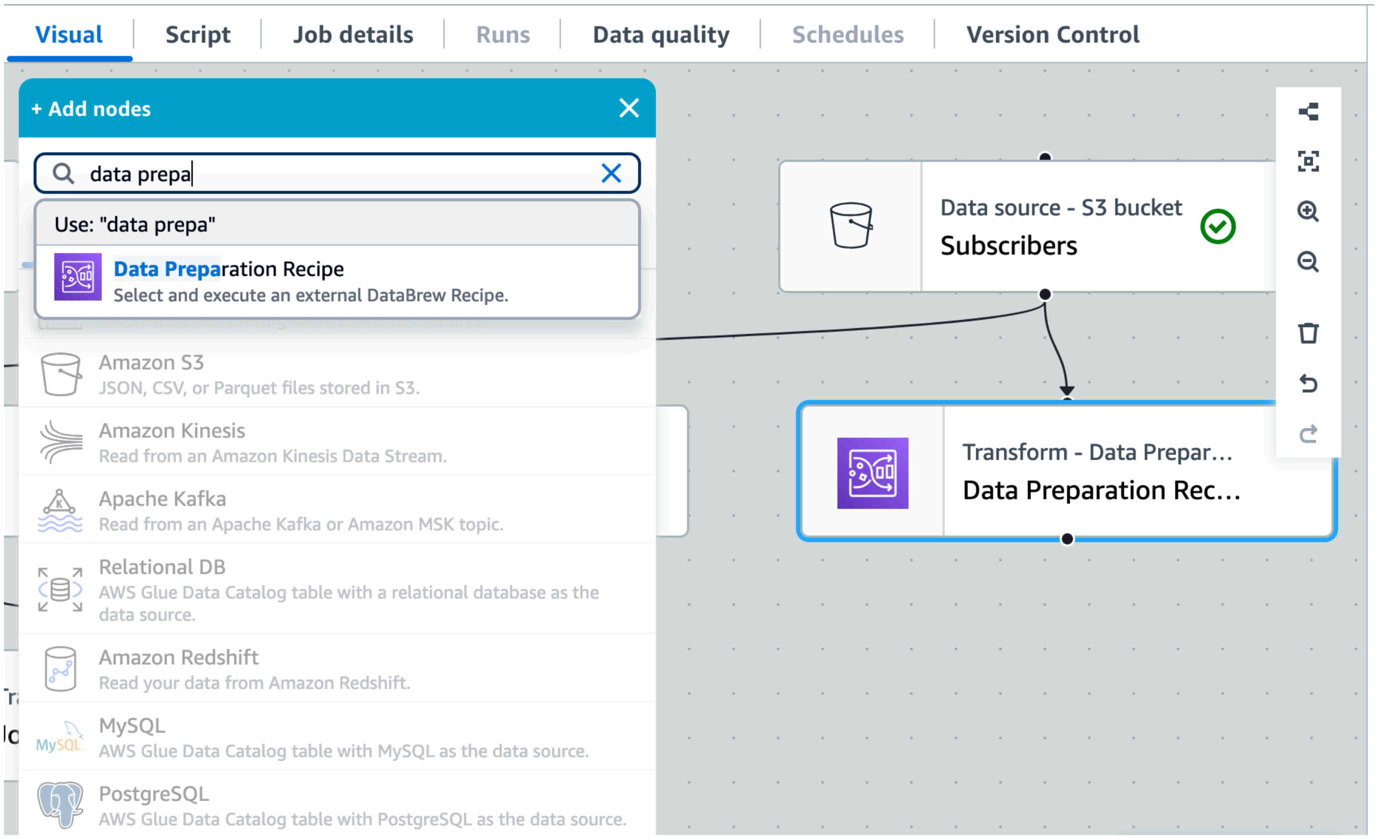
```
"errorCode": "AccessDenied"
"errorMessage": "User: arn:aws:sts::account_id:assumed-role/AWSGlueServiceRole is not
authorized to perform: iam:PassRole on resource: arn:aws:iam::account_id:role/service-
role/AWSGlueServiceRole
because no identity-based policy allows the iam:PassRole action"
```

Importing an AWS Glue DataBrew recipe

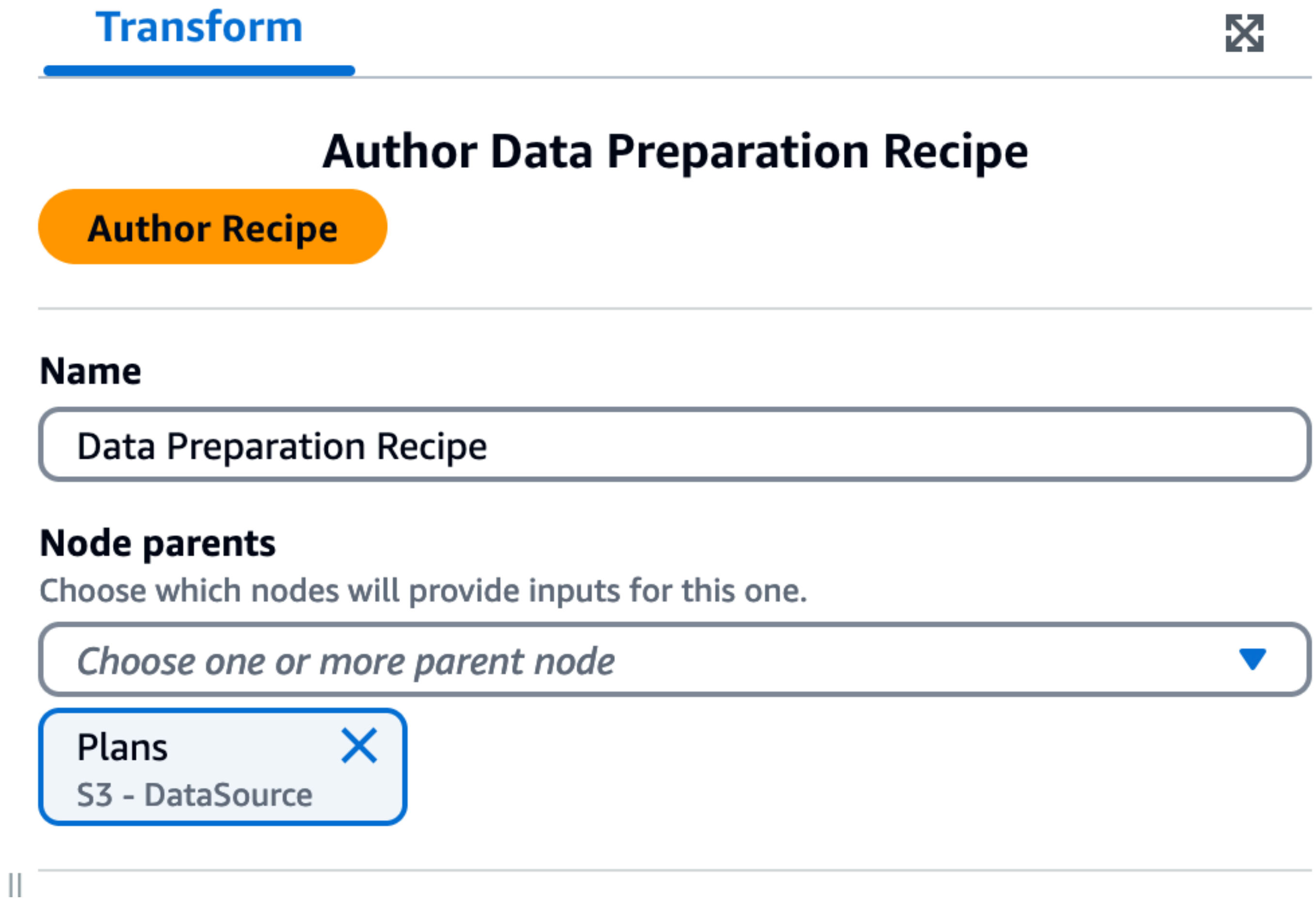
To import an AWS Glue DataBrew recipe and use in AWS Glue Studio:

If you have an existing **Data Preparation Recipe** node and you want to edit the recipe steps directly in AWS Glue Studio, you will have to import the recipe steps into your AWS Glue Studio job.

1. Start a AWS Glue job in AWS Glue Studio with a datasource.
2. Add the **Data Preparation Recipe** node to the job canvas.


Visual
Script
Job details
Runs
Data quality
Schedules
Version Control
+ Add nodes
data prepa
Use: "data prepa"
Data Preparation Recipe
Select and execute an external DataBrew Recipe.
Amazon S3
JSON, CSV, or Parquet files stored in S3.
Amazon Kinesis
Read from an Amazon Kinesis Data Stream.
Apache Kafka
Read from an Apache Kafka or Amazon MSK topic.
Relational DB
AWS Glue Data Catalog table with a relational database as the data source.
Amazon Redshift
Read your data from Amazon Redshift.
MySQL
AWS Glue Data Catalog table with MySQL as the data source.
PostgreSQL
AWS Glue Data Catalog table with PostgreSQL as the data source.
Data source - S3 bucket
Subscribers
Transform - Data Prepar…
Data Preparation Rec…

3. In the Transform panel, enter a name for your recipe.
4. Choose one or more parent nodes by selecting the available nodes on the canvas from the drop-down list.
5. Choose **Author Recipe**. If **Author Recipe** is grey it is unavailable until node parents have been selected and a data preview session has finished.



Author Data Preparation Recipe

Name

Data Preparation Recipe

Node parents

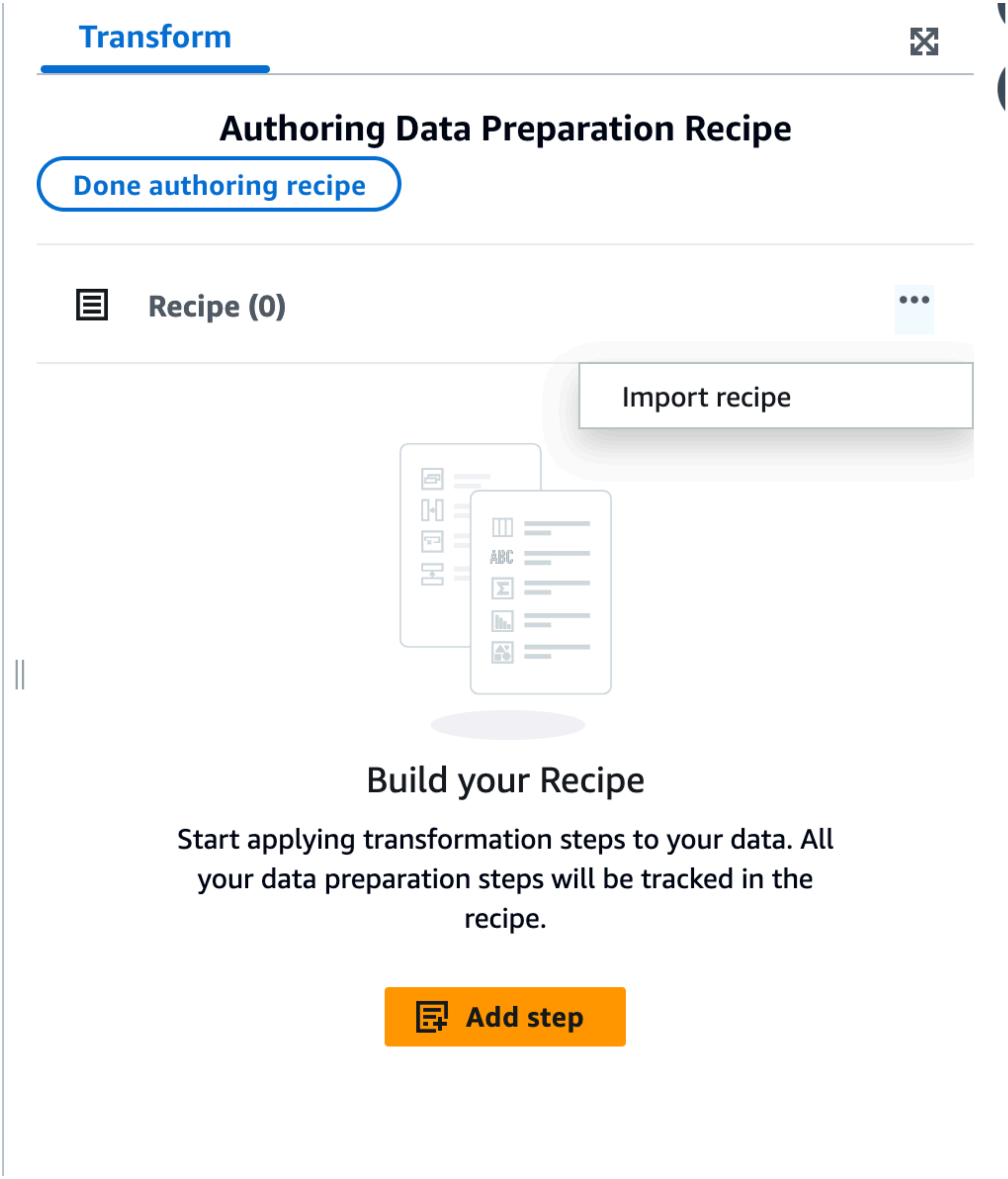
Choose which nodes will provide inputs for this one.

Choose one or more parent node

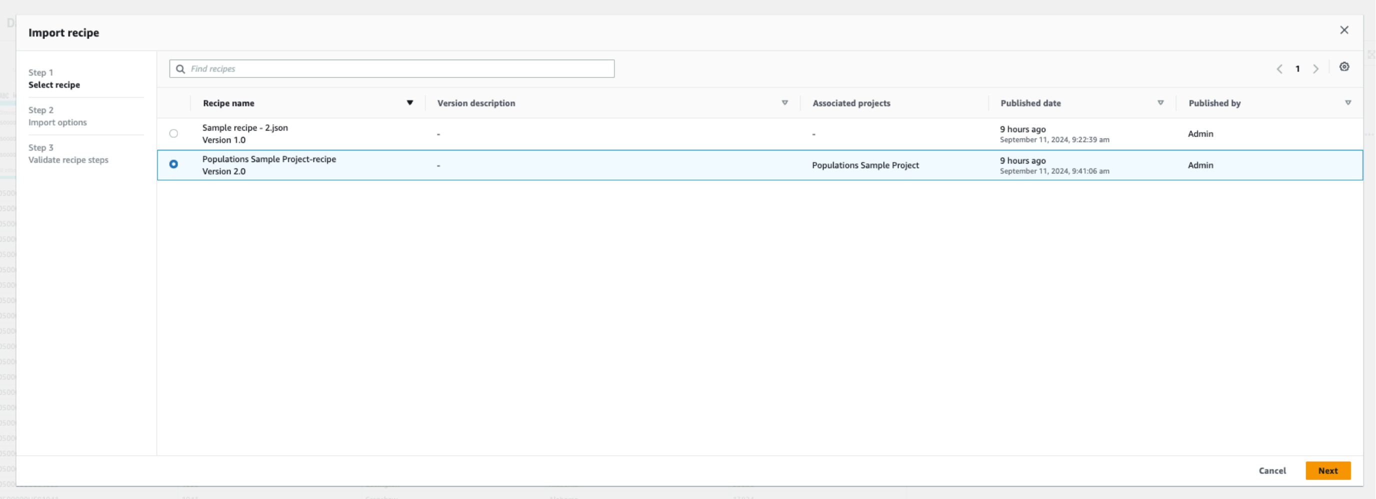
Plans
S3 - DataSource

6. The data frame loads and shows you detailed information about your source data.

   Select the **more actions** icon and choose **Import recipe**.


Transform
Authoring Data Preparation Recipe
Done authoring recipe
Recipe (0)
•••
Import recipe
ABC
Build your Recipe
Start applying transformation steps to your data. All your data preparation steps will be tracked in the recipe.
Add step

7. Use the Import recipe wizard to complete the steps. In step 1, search for your recipe, select it, and choose **Next**.



Import recipe

Step 1
Select recipe

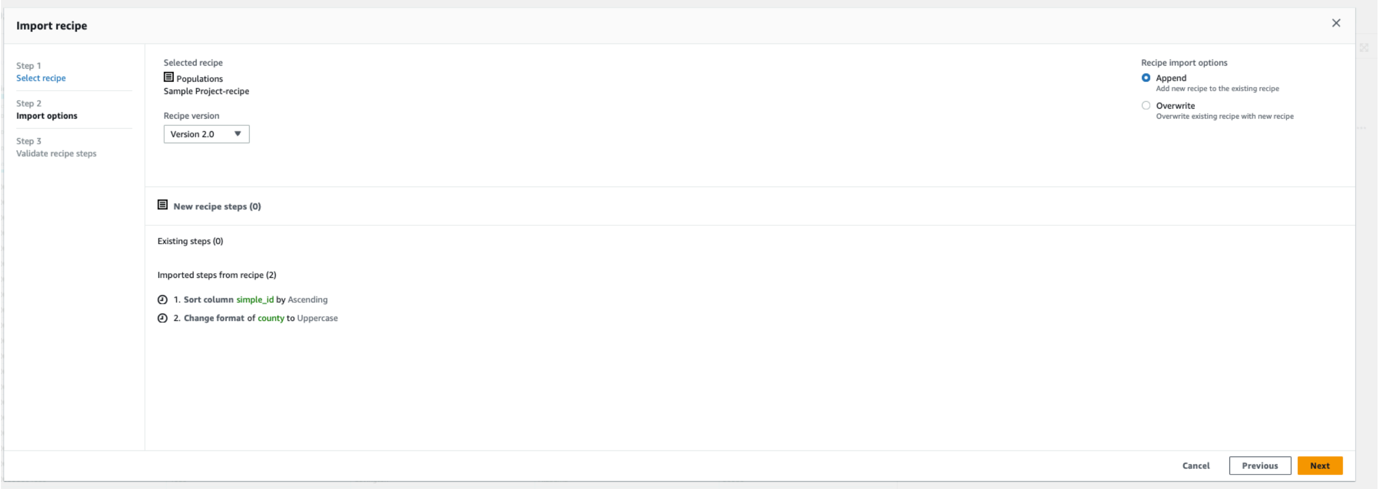
Step 2
Import options

Step 3
Validate recipe steps

Find recipes

	Recipe name	Version description	Associated projects	Published date	Published by
○	Sample recipe - 2.json Version 1.0	-	-	9 hours ago September 11, 2024, 9:22:39 am	Admin
●	Populations Sample Project-recipe Version 2.0	-	Populations Sample Project	9 hours ago September 11, 2024, 9:41:06 am	Admin

8. In step 2, choose your import options. You can choose to Append a new recipe to an existing recipe or Overwrite an existing recipe. Choose **Next**.



Import recipe

Step 1
Select recipe

Step 2
Import options

Step 3
Validate recipe steps

Selected recipe
Populations
Sample Project-recipe

Recipe version
Version 2.0

Recipe import options
● Append
Add new recipe to the existing recipe
○ Overwrite
Overwrite existing recipe with new recipe

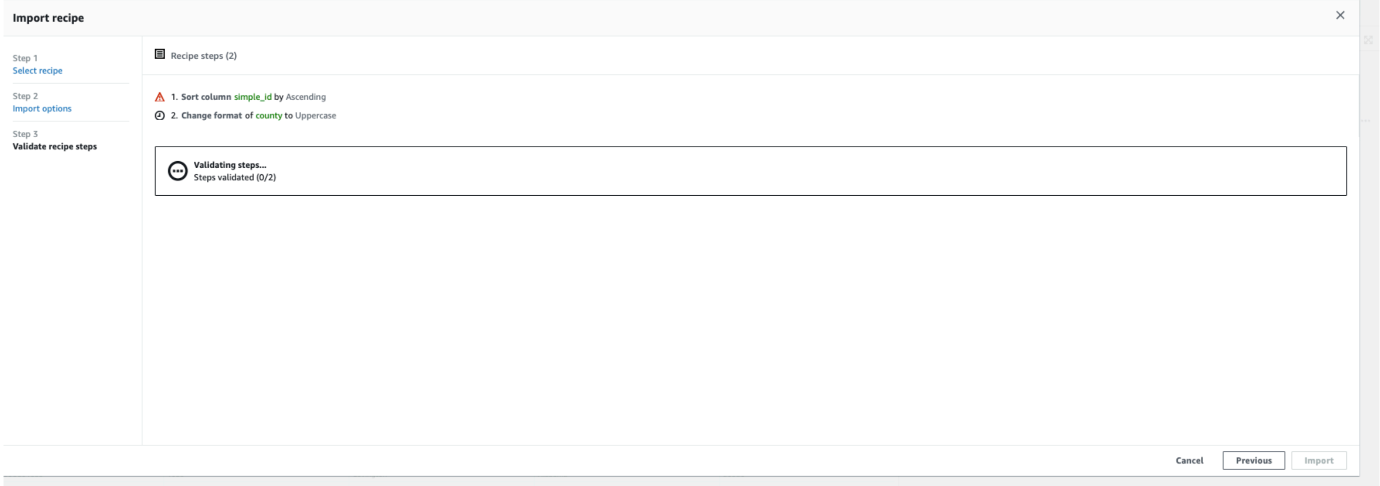
New recipe steps (0)

Existing steps (0)

Imported steps from recipe (2)

1. Sort column simple_id by Ascending
2. Change format of county to Uppercase

9. In step 3, validate the recipe steps. Once you import your AWS Glue DataBrew recipe, you can edit this recipe directly in AWS Glue Studio.



Import recipe

Step 1
Select recipe

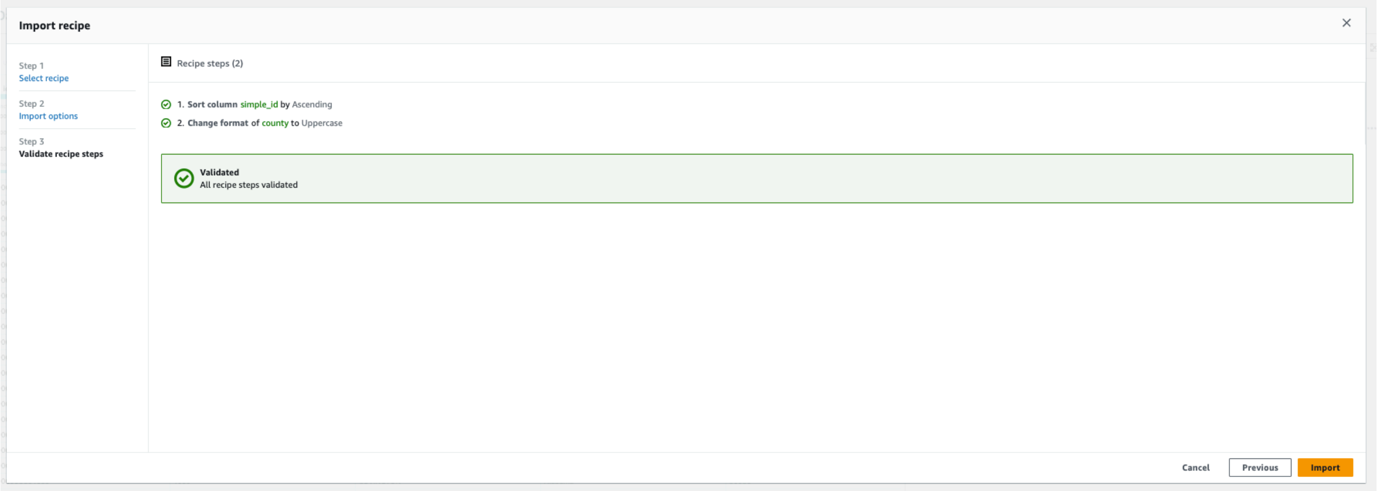
Step 2
Import options

Step 3
Validate recipe steps

Recipe steps (2)

1. Sort column simple_id by Ascending
2. Change format of county to Uppercase

Validating steps...
Steps validated (0/2)


Import recipe
Step 1
Select recipe
Step 2
Import options
Step 3
Validate recipe steps
Recipe steps (2)
Cancel
Previous
Import

1. Sort column simple_id by Ascending
2. Change format of county to Uppercase

Validated
All recipe steps validated

10. After this, the steps will be imported as part of your AWS Glue job. Make necessary configuration changes in the **Job details** tab, like naming your job and adjusting allocated capacity as needed. Choose **Save** to save your job and recipe.

> **Note**
>
> JOIN, UNION, GROUP_BY, PIVOT, UNPIVOT, TRANSPOSE are not supported for recipe import, nor will they be available in recipe authoring mode.

11. Optionally, you can finish authoring the job by adding other transformations nodes as needed and add Data target node(s).

    If you reorder steps after you import a recipe, AWS Glue performs validation on those steps. For example, if you renamed and then deleted a column, and you moved the delete step on top, then the rename step would be invalid. You can then edit the steps to fix the validation error.

Migrating from AWS Glue DataBrew to AWS Glue Studio

If you have recipes in AWS Glue DataBrew, use the following checklist to migrate your recipes to AWS Glue Studio.

If you want to	Then do this
Allow users to retrieve AWS Glue DataBrew recipes, recipe versions, and recipe descriptions.	Add IAM permissions to a policy that allows your role to access the necessary actions. See IAM permissions for AWS Glue DataBrew.
Import an existing AWS Glue DataBrew recipe into AWS Glue Studio.	Follow the steps in Importing an AWS Glue DataBrew recipe .
Import a recipe with JOIN and UNION.	Recipes with UNION and JOIN transforms are not supported. Use the **Join** and **Union** transforms in AWS Glue Studio before or after a Data Preparation Recipe node.

Using Change Schema to remap data property keys

A *Change Schema* transform remaps the source data property keys into the desired configured for the target data. In a Change Schema transform node, you can:

- Change the name of multiple data property keys.
- Change the data type of the data property keys, if the new data type is supported and there is a transformation path between the two data types.
- Choose a subset of data property keys by indicating which data property keys you want to drop.

You can also add additional *Change Schema* nodes to the job diagram as needed – for example, to modify additional data sources or following a *Join* transform.

Using Change Schema with decimal datatype

When using the **Change Schema** transform with decimal datatype, the **Change Schema** transform modifies the precision to the default value of (10,2). To modify this and set the precision for your use case, you can use the **SQL Query** transform and cast the columns with a specific precision.

For example, if you have an input column named "DecimalCol" of type Decimal, and you want to remap it to an output column named "OutputDecimalCol" with a specific precision of (18,6), you would:

1. Add a subsequent **SQL Query** transform after the **Change Schema** transform.

2. In the **SQL Query** transform, use an SQL query to cast the remapped column to the desired precision. The SQL query would look like this:

   ```
   SELECT col1, col2, CAST(DecimalCol AS DECIMAL(18,6)) AS OutputDecimalCol
   FROM __THIS__
   ```

   In the above SQL query:

   - `col1` and `col2` are other columns in your data that you want to pass through without modification.
   - `DecimalCol` is the original column name from the input data.
   - `CAST(DecimalCol AS DECIMAL(18,6))` casts the `DecimalCol` to a Decimal type with a precision of 18 digits and 6 decimal places.
   - `AS OutputDecimalCol` renames the casted column to `OutputDecimalCol`.

By using the **SQL Query** transform, you can override the default precision set by the **Change Schema** transform and explicitly cast the Decimal columns to the desired precision. This approach allows you to leverage the **Change Schema** transform for renaming and restructuring your data while handling the precision requirements for Decimal columns through the subsequent **SQL Query** transformation.

Adding a Change Schema transform to your job

> **Note**
>
> The **Change Schema** transform is not case-sensitive.

To add a Change Schema transform node to your job diagram

1. (Optional) Open the Resource panel and then choose **Change Schema** to add a new transform to your job diagram, if needed.
2. In the node properties panel, enter a name for the node in the job diagram. If a node parent isn't already selected, choose a node from the **Node parents** list to use as the input source for the transform.
3. Choose the **Transform** tab in the node properties panel.
4. Modify the input schema:

- To rename a data property key, enter the new name of the key in the **Target key** field.
- To change the data type for a data property key, choose the new data type for the key from the **Data type** list.
- To remove a data property key from the target schema, choose the **Drop** check box for that key.

5. (Optional) After configuring the transform node properties, you can view the modified schema for your data by choosing the **Output schema** tab in the node details panel. The first time you choose this tab for any node in your job, you are prompted to provide an IAM role to access the data. If you have not specified an IAM role on the **Job details** tab, you are prompted to enter an IAM role here.
6. (Optional) After configuring the node properties and transform properties, you can preview the modified dataset by choosing the **Data preview** tab in the node details panel. The first time you choose this tab for any node in your job, you are prompted to provide an IAM role to access the data. There is a cost associated with using this feature, and billing starts as soon as you provide an IAM role.

Using Drop Duplicates

The Drop Duplicates transform removes rows from your data source by giving you two options. You can choose to remove the duplicate row that are completely the same, or you can choose to choose the fields to match and remove only those rows based on your chosen fields.

For example, in this data set, you have duplicate rows where all the values in some of the rows are exactly the same as another row, and some of the values in rows are the same or different.

Row	Name	Email	Age	State	Note
1	Joy	joy@gmail	33	NY	
2	Tim	tim@gmail	45	OH	
3	Rose	rose@gmail	23	NJ	
4	Tim	tim@gmail	42	OH	
5	Rose	rose@gmail	23	NJ	

Row	Name	Email	Age	State	Note
6	Tim	tim@gmail	42	OH	this is a duplicate row and matches completely on all values as row #4
7	Rose	rose@gmail	23	NJ	This is a duplicate row and matches completely on all values as row #5

If you choose to match entire rows, rows 6 and 7 will be removed from the data set. The data set is now:

Row	Name	Email	Age	State
1	Joy	joy@gmail	33	NY
2	Tim	tim@gmail	45	OH
3	Rose	rose@gmail	23	NJ
4	Tim	tim@gmail	42	OH
5	Rose	rose@gmail	23	NJ

If you chose to specify keys, you can choose to remove rows that match on ‘name’ and ‘email’. This gives you finer control of what is a ‘duplicate row’ for your data set. By specifying ‘name’ and ‘email’, the data set is now:

Row	Name	Email	Age	State
1	Joy	joy@gmail	33	NY
2	Tim	tim@gmail	45	OH
3	Rose	rose@gmail	23	NJ

Some things to keep in mind:

- In order for rows to be recognized as a duplicate, values are case sensitive. all values in rows need to have the same casing - this applies to either option you choose (Match entire rows or Specify keys).
- All values are read in as strings.
- The **Drop Duplicates** transform utilizes the Spark dropDuplicates command.
- When using the **Drop Duplicates** transform, the first row is kept and other rows are dropped.
- The **Drop Duplicates** transform does not change the schema of the dataframe. If you choose to specify keys, all fields are kept in the resulting dataframe.

Using SelectFields to remove most data property keys

You can create a subset of data property keys from the dataset using the *SelectFields* transform. You indicate which data property keys you want to keep and the rest are removed from the dataset.

> **Note**
>
> The *SelectFields* transform is case sensitive. Use *ApplyMapping* if you need a case-insensitive way to select fields.

To add a SelectFields transform node to your job diagram

1. (Optional) Open the Resource panel, and then choose **SelectFields** to add a new transform to your job diagram, if needed.

2. On the **Node properties** tab, enter a name for the node in the job diagram. If a node parent is not already selected, choose a node from the **Node parents** list to use as the input source for the transform.
3. Choose the **Transform** tab in the node details panel.
4. Under the heading **SelectFields**, choose the data property keys in the dataset that you want to keep. Any data property keys not selected are dropped from the dataset.

   You can also choose the check box next to the column heading **Field** to automatically choose all the data property keys in the dataset. Then you can deselect individual data property keys to remove them from the dataset.
5. (Optional) After configuring the transform node properties, you can view the modified schema for your data by choosing the **Output schema** tab in the node details panel. The first time you choose this tab for any node in your job, you are prompted to provide an IAM role to access the data. If you have not specified an IAM role on the **Job details** tab, you are prompted to enter an IAM role here.
6. (Optional) After configuring the node properties and transform properties, you can preview the modified dataset by choosing the **Data preview** tab in the node details panel. The first time you choose this tab for any node in your job, you are prompted to provide an IAM role to access the data. There is a cost associated with using this feature, and billing starts as soon as you provide an IAM role.

Using DropFields to keep most data property keys

You can create a subset of data property keys from the dataset using the *DropFields* transform. You indicate which data property keys you want to remove from the dataset and the rest of the keys are retained.

> **Note**
>
> The *DropFields* transform is case sensitive. Use *Change Schema* if you need a case-insensitive way to select fields.

To add a DropFields transform node to your job diagram

1. (Optional) Open the Resource panel and then choose **DropFields** to add a new transform to your job diagram, if needed.

2. On the **Node properties** tab, enter a name for the node in the job diagram. If a node parent is not already selected, then choose a node from the **Node parents** list to use as the input source for the transform.
3. Choose the **Transform** tab in the node details panel.
4. Under the heading **DropFields**, choose the data property keys to drop from the data source.

   You can also choose the check box next to the column heading **Field** to automatically choose all the data property keys in the dataset. Then you can deselect individual data property keys so they are retained in the dataset.
5. (Optional) After configuring the transform node properties, you can view the modified schema for your data by choosing the **Output schema** tab in the node details panel. The first time you choose this tab for any node in your job, you are prompted to provide an IAM role to access the data. If you have not specified an IAM role on the **Job details** tab, you are prompted to enter an IAM role here.
6. (Optional) After configuring the node properties and transform properties, you can preview the modified dataset by choosing the **Data preview** tab in the node details panel. The first time you choose this tab for any node in your job, you are prompted to provide an IAM role to access the data. There is a cost associated with using this feature, and billing starts as soon as you provide an IAM role.

Renaming a field in the dataset

You can use the *RenameField* transform to change the name for an individual property key in the dataset.

> **Note**
>
> The *RenameField* transform is case sensitive. Use *ApplyMapping* if you need a case-insensitive transform.

> **Tip**
>
> If you use the *Change Schema* transform, you can rename multiple data property keys in the dataset with a single transform.

To add a RenameField transform node to your job diagram

1. (Optional) Open the Resource panel and then choose **RenameField** to add a new transform to your job diagram, if needed.
2. On the **Node properties** tab, enter a name for the node in the job diagram. If a node parent is not already selected, then choose a node from the **Node parents** list to use as the input source for the transform.
3. Choose the **Transform** tab.
4. Under the heading **Data field**, choose a property key from the source data and then enter a new name in the **New field name** field.
5. (Optional) After configuring the transform node properties, you can view the modified schema for your data by choosing the **Output schema** tab in the node details panel. The first time you choose this tab for any node in your job, you are prompted to provide an IAM role to access the data. If you have not specified an IAM role on the **Job details** tab, you are prompted to enter an IAM role here.
6. (Optional) After configuring the node properties and transform properties, you can preview the modified dataset by choosing the **Data preview** tab in the node details panel. The first time you choose this tab for any node in your job, you are prompted to provide an IAM role to access the data. There is a cost associated with using this feature, and billing starts as soon as you provide an IAM role.

Using Spigot to sample your dataset

To test the transformations performed by your job, you might want to get a sample of the data to check that the transformation works as intended. The *Spigot* transform writes a subset of records from the dataset to a JSON file in an Amazon S3 bucket. The data sampling method can be either a specific number of records from the beginning of the file or a probability factor used to pick records.

To add a Spigot transform node to your job diagram

1. (Optional) Open the Resource panel and then choose **Spigot** to add a new transform to your job diagram, if needed.
2. On the **Node properties** tab, enter a name for the node in the job diagram. If a node parent is not already selected, then choose a node from the **Node parents** list to use as the input source for the transform.

3. Choose the **Transform** tab in the node details panel.
4. Enter an Amazon S3 path or choose **Browse S3** to choose a location in Amazon S3. This is the location where the job writes the JSON file that contains the data sample.
5. Enter information for the sampling method. You can specify a value for **Number of records** to write starting from the beginning of the dataset and a **Probability threshold** (entered as a decimal value with a maximum value of 1) of picking any given record.

   For example, to write the first 50 records from the dataset, you would set **Number of records** to 50 and **Probability threshold** to 1 (100%).

Joining datasets

The *Join* transform allows you to combine two datasets into one. You specify the key names in the schema of each dataset to compare. The output `DynamicFrame` contains rows where keys meet the join condition. The rows in each dataset that meet the join condition are combined into a single row in the output `DynamicFrame` that contains all the columns found in either dataset.

To add a Join transform node to your job diagram

1. If there is only one data source available, you must add a new data source node to the job diagram.
2. Choose one of the source nodes for the join. Open the Resource panel and then choose **Join** to add a new transform to your job diagram.
3. On the **Node properties** tab, enter a name for the node in the job diagram.
4. In the **Node properties** tab, under the heading **Node parents**, add a parent node so that there are two datasets providing inputs for the join. The parent can be a data source node or a transform node.

   > **Note**
   >
   > A join can have only two parent nodes.

5. Choose the **Transform** tab.

   If you see a message indicating that there are conflicting key names, you can either:

   - Choose **Resolve it** to automatically add an *ApplyMapping* transform node to your job diagram. The ApplyMapping node adds a prefix to any keys in the dataset that have the same name as a key in the other dataset. For example, if you use the default value of **`right`**, then any keys in the right dataset that have the same name as a key in the left dataset will be renamed to `(right)`*`key name`*.

  - Manually add a transform node earlier in the job diagram to remove or rename the conflicting keys.

6. Choose the type of join in the **Join type** list.

   - **Inner join**: Returns a row with columns from both datasets for every match based on the join condition. Rows that don't satisfy the join condition aren't returned.
   - **Left join**: All rows from the left dataset and only the rows from the right dataset that satisfy the join condition.
   - **Right join**: All rows from the right dataset and only the rows from the left dataset that satisfy the join condition.
   - **Outer join**: All rows from both datasets.
   - **Left semi join**: All rows from the left dataset that have a match in the right dataset based on the join condition.
   - **Left anti join**: All rows in the left dataset that don't have a match in the right dataset based on join condition.

7. On the **Transform** tab, under the heading **Join conditions**, choose **Add condition**. Choose a property key from each dataset to compare. Property keys on the left side of the comparison operator are referred to as the left dataset and property keys on the right are referred to as the right dataset.

   For more complex join conditions, you can add additional matching keys by choosing **Add condition** more than once. If you accidentally add a condition, you can choose the delete icon (🗑) to remove it.

8. (Optional) After configuring the transform node properties, you can view the modified schema for your data by choosing the **Output schema** tab in the node details panel. The first time you choose this tab for any node in your job, you are prompted to provide an IAM role to access the data. If you have not specified an IAM role on the **Job details** tab, you are prompted to enter an IAM role here.

9. (Optional) After configuring the node properties and transform properties, you can preview the modified dataset by choosing the **Data preview** tab in the node details panel. The first time you choose this tab for any node in your job, you are prompted to provide an IAM role to

access the data. There is a cost associated with using this feature, and billing starts as soon as you provide an IAM role.

For an example of the join output schema, consider a join between two datasets with the following property keys:

```
Left: {id, dept, hire_date, salary, employment_status}
Right: {id, first_name, last_name, hire_date, title}
```

The join is configured to match on the `id` and `hire_date` keys using the = comparison operator.

Because both datasets contain `id` and `hire_date` keys, you chose **Resolve it** to automatically add the prefix **`right`** to the keys in the right dataset.

The keys in the output schema would be:

```
{id, dept, hire_date, salary, employment_status,
(right)id, first_name, last_name, (right)hire_date, title}
```

Using Union to combine rows

You use the Union transform node when you want to combine rows from more than one data source that have the same schema.

There are to types of Union transformations:

1. ALL – when applying ALL, the resulting union does not remove duplicate rows.
2. DISTINCT – when applying DISTINCT, the resulting union removes duplicate rows.

Unions vs. Joins

You use Union to combine rows. You use Join to combine columns.

Using the Union transform in the Visual ETL canvas

1. Add more than one data source to perform a union transform. To add a data source, open the Resource Panel, then choose the data source from the Sources tab. Before using the Union transformation, you must ensure that all data sources involved in the union have the same schema and structure.

2. When you have at least two data sources that you want to combine using the Union transform, create the Union transform by adding it to the canvas. Open the Resource Panel on the canvas and search for 'Union'. You can also choose the Transforms tab in the Resource Panel and scroll down until you find the Union transform, then choose **Union**.
3. Select the Union node on the job canvas. In the Node properties window, choose the parent nodes to connect to the Union transform.
4. AWS Glue checks for compatibility to make sure that the Union transform can be applied to all data sources. If the schema for the data sources are the same, the operation will be allowed. If the data sources do not have the same schema, an invalid error message is displayed: “The input schemas of this union are not the same. Consider using ApplyMapping to match the schemas.” To fix this, choose Use **ApplyMapping**.
5. Choose the Union type.

    1. All – By default, the All Union type is selected; this will result in duplicate rows if there are any in the data combination.
    2. Distinct – Choose Distinct if you want duplicate rows to be removed from the resulting data combination.

Using SplitFields to split a dataset into two

The *SplitFields* transform allows you to choose some of the data property keys in the input dataset and put them into one dataset and the unselected keys into a separate dataset. The output from this transform is a collection of `DynamicFrames`.

> **Note**
>
> You must use a *SelectFromCollection* transform to convert the collection of `DynamicFrames` into a single `DynamicFrame` before you can send the output to a target location.

The *SplitFields* transform is case sensitive. Add an *ApplyMapping* transform as a parent node if you need case-insensitive property key names.

To add a SplitFields transform node to your job diagram

1. (Optional) Open the Resource panel and then choose **SplitFields** to add a new transform to your job diagram, if needed.
2. On the **Node properties** tab, enter a name for the node in the job diagram. If a node parent is not already selected, then choose a node from the **Node parents** list to use as the input source for the transform.
3. Choose the **Transform** tab.
4. Choose which property keys you want to put into the first dataset. The keys that you do not choose are placed in the second dataset.
5. (Optional) After configuring the transform node properties, you can view the modified schema for your data by choosing the **Output schema** tab in the node details panel. The first time you choose this tab for any node in your job, you are prompted to provide an IAM role to access the data. If you have not specified an IAM role on the **Job details** tab, you are prompted to enter an IAM role here.
6. (Optional) After configuring the node properties and transform properties, you can preview the modified dataset by choosing the **Data preview** tab in the node details panel. The first time you choose this tab for any node in your job, you are prompted to provide an IAM role to access the data. There is a cost associated with using this feature, and billing starts as soon as you provide an IAM role.
7. Configure a *SelectFromCollection* transform node to process the resulting datasets.

Overview of *SelectFromCollection* transform

Certain transforms have multiple datasets as their output instead of a single dataset, for example, *SplitFields*. The *SelectFromCollection* transform selects one dataset (`DynamicFrame`) from a collection of datasets (an array of `DynamicFrames`). The output for the transform is the selected `DynamicFrame`.

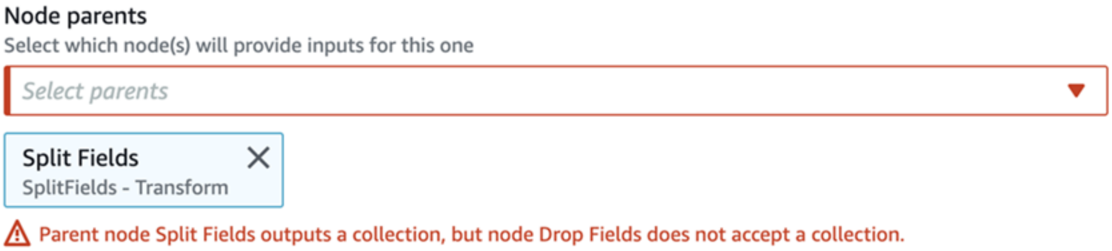
You must use this transform after you use a transform that creates a collection of `DynamicFrames`, such as:

- Custom code transforms
- *SplitFields*

If you don't add a *SelectFromCollection* transform node to your job diagram after any of these transforms, you will get an error for your job.

The parent node for this transform must be a node that returns a collection of `DynamicFrames`. If you choose a parent for this transform node that returns a single `DynamicFrame`, such as a *Join* transform, your job returns an error.

Similarly, if you use a *SelectFromCollection* node in your job diagram as the parent for a transform that expects a single `DynamicFrame` as input, your job returns an error.



Node parents

Select which node(s) will provide inputs for this one

Select parents

Split Fields

SplitFields - Transform

Parent node Split Fields outputs a collection, but node Drop Fields does not accept a collection.

Using SelectFromCollection to choose which dataset to keep

Use the *SelectFromCollection* transform to convert a collection of `DynamicFrames` into a single `DynamicFrame`.

To add a SelectFromCollection transform node to your job diagram

1. (Optional) Open the Resource panel and then choose **SelectFromCollection** to add a new transform to your job diagram, if needed.
2. On the **Node properties** tab, enter a name for the node in the job diagram. If a node parent is not already selected, then choose a node from the **Node parents** list to use as the input source for the transform.
3. Choose the **Transform** tab.
4. Under the heading **Frame index**, choose the array index number that corresponds to the `DynamicFrame` you want to select from the collection of `DynamicFrames`.

   For example, if the parent node for this transform is a *SplitFields* transform, on the **Output schema** tab of that node you can see the schema for each `DynamicFrame`. If you want to keep the `DynamicFrame` associated with the schema for **Output 2**, you would select **1** for the value of **Frame index**, which is the second value in the list.

   Only the `DynamicFrame` that you choose is included in the output.

5. (Optional) After configuring the transform node properties, you can view the modified schema for your data by choosing the **Output schema** tab in the node details panel. The first time you choose this tab for any node in your job, you are prompted to provide an IAM role to access the data. If you have not specified an IAM role on the **Job details** tab, you are prompted to enter an IAM role here.
6. (Optional) After configuring the node properties and transform properties, you can preview the modified dataset by choosing the **Data preview** tab in the node details panel. The first time you choose this tab for any node in your job, you are prompted to provide an IAM role to access the data. There is a cost associated with using this feature, and billing starts as soon as you provide an IAM role.

Find and fill missing values in a dataset

You can use the *FillMissingValues* transform to locate records in the dataset that have missing values and add a new field with a value determined by imputation. The input data set is used to train the machine learning (ML) model that determines what the missing value should be. If you use incremental data sets, then each incremental set is used as the training data for the ML model, so the results might not be as accurate.

To use a FillMissingValues transform node in your job diagram

1. (Optional) Open the Resource panel and then choose **FillMissingValues** to add a new transform to your job diagram, if needed.
2. On the **Node properties** tab, enter a name for the node in the job diagram. If a node parent isn't already selected, choose a node from the **Node parents** list to use as the input source for the transform.
3. Choose the **Transform** tab.
4. For **Data field**, choose the column or field name from the source data that you want to analyze for missing values.
5. (Optional) In the **New field name** field, enter a name for the field added to each record that will hold the estimated replacement value for the analyzed field. If the analyzed field doesn't have a missing value, the value in the analyzed field is copied into the new field.

   If you don't specify a name for the new field, the default name is the name of the analyzed column with `_filled` appended. For example, if you enter **Age** for **Data field** and don't specify a value for **New field name**, a new field named **`Age_filled`** is added to each record.

6. (Optional) After configuring the transform node properties, you can view the modified schema for your data by choosing the **Output schema** tab in the node details panel. The first time you choose this tab for any node in your job, you are prompted to provide an IAM role to access the data. If you have not specified an IAM role on the **Job details** tab, you are prompted to enter an IAM role here.
7. (Optional) After configuring the node properties and transform properties, you can preview the modified dataset by choosing the **Data preview** tab in the node details panel. The first time you choose this tab for any node in your job, you are prompted to provide an IAM role to access the data. There is a cost associated with using this feature, and billing starts as soon as you provide an IAM role.

Filtering keys within a dataset

Use the *Filter* transform to create a new dataset by filtering records from the input dataset based on a regular expression. Rows that don't satisfy the filter condition are removed from the output.

- For string data types, you can filter rows where the key value matches a specified string.
- For numeric data types, you can filter rows by comparing the key value to a specified value using the comparison operators <, >, =, !=, <=, and >=.

If you specify multiple filter conditions, the results are combined using an `AND` operator by default, but you can choose `OR` instead.

The *Filter* transform is case sensitive. Add an *ApplyMapping* transform as a parent node if you need case-insensitive property key names.

To add a Filter transform node to your job diagram

1. (Optional) Open the Resource panel and then choose **Filter** to add a new transform to your job diagram, if needed.
2. On the **Node properties** tab, enter a name for the node in the job diagram. If a node parent isn't already selected, then choose a node from the **Node parents** list to use as the input source for the transform.
3. Choose the **Transform** tab.
4. Choose either **Global AND** or **Global OR**. This determines how multiple filter conditions are combined. All conditions are combined using either `AND` or `OR` operations. If you have only a single filter conditions, then you can choose either one.

5. Choose the **Add condition** button in the **Filter condition** section to add a filter condition.

   In the **Key** field, choose a property key name from the dataset. In the **Operation** field, choose the comparison operator. In the **Value** field, enter the comparison value. Here are some examples of filter conditions:

   - `year >= 2018`
   - `State matches 'CA*'`

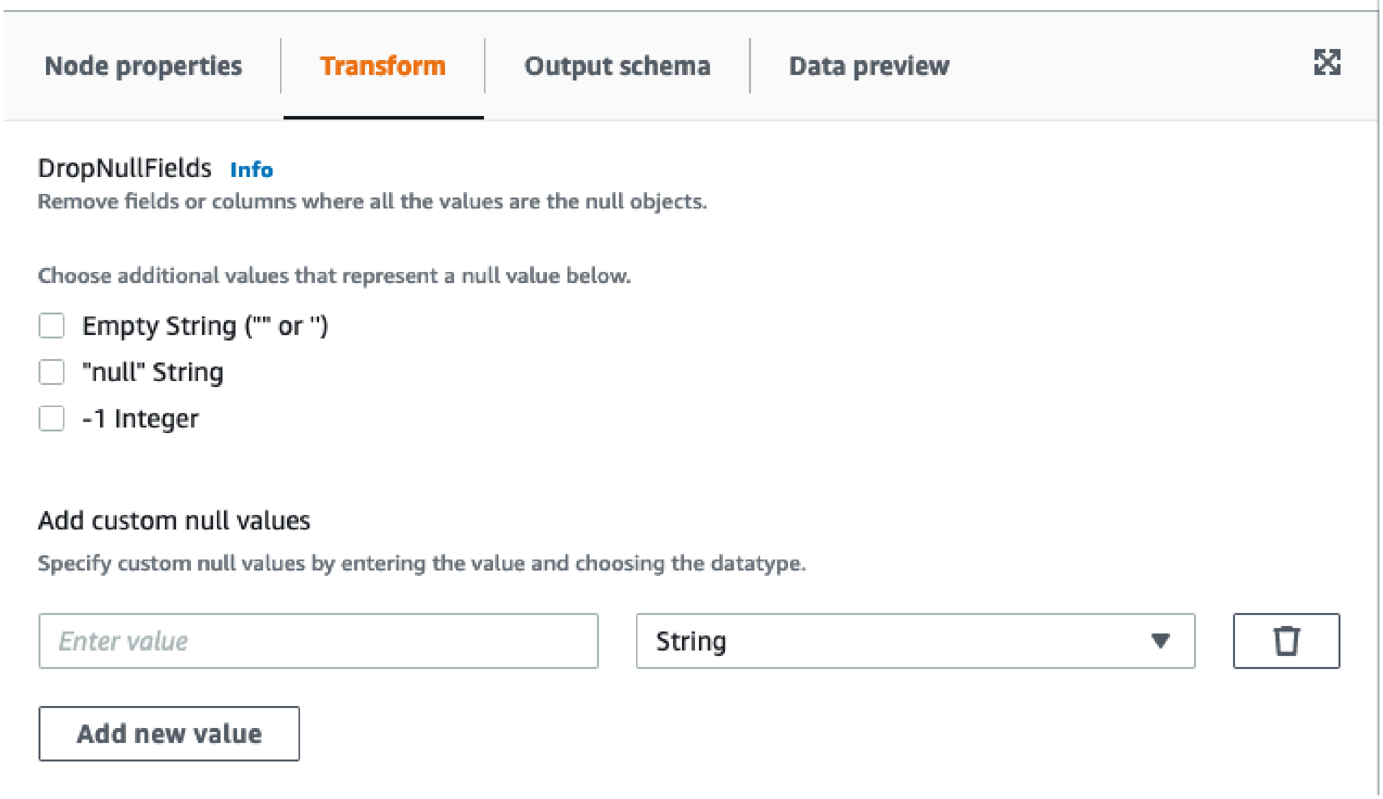
   When you filter on string values, make sure that the comparison value uses a regular expression format that matches the script language selected in the job properties (Python or Scala).
6. Add additional filter conditions, as needed.
7. (Optional) After configuring the transform node properties, you can view the modified schema for your data by choosing the **Output schema** tab in the node details panel. The first time you choose this tab for any node in your job, you are prompted to provide an IAM role to access the data. If you have not specified an IAM role on the **Job details** tab, you are prompted to enter an IAM role here.
8. (Optional) After configuring the node properties and transform properties, you can preview the modified dataset by choosing the **Data preview** tab in the node details panel. The first time you choose this tab for any node in your job, you are prompted to provide an IAM role to access the data. There is a cost associated with using this feature, and billing starts as soon as you provide an IAM role.

Using DropNullFields to remove fields with null values

Use the *DropNullFields* transform to remove fields from the dataset if all values in the field are ‘null’. By default, AWS Glue Studio will recognize null objects, but some values such as empty strings, strings that are “null”, -1 integers or other placeholders such as zeros, are not automatically recognized as nulls.

To use the DropNullFields

1. Add a DropNullFields node to the job diagram.
2. On the **Node properties** tab, choose additional values that represent a null value. You can choose to select none or all of the values:



DropNullFields Info

Remove fields or columns where all the values are the null objects.

Choose additional values that represent a null value below.

☐ Empty String ("" or '')

☐ "null" String

☐ -1 Integer

Add custom null values

Specify custom null values by entering the value and choosing the datatype.

- Empty String ("" or '') - fields that contain empty strings will be removed
- "null string" - fields that contain the string with the word 'null' will be removed
- -1 integer - fields that contain a -1 (negative one) integer will be removed

3. If needed, you can also specify custom null values. These are null values that may be unique to your dataset. To add a custom null value, choose **Add new value**.
4. Enter the custom null value. For example, this can zero, or any value that is being used to represent a null in the dataset.
5. Choose the data type in the drop-down field. Data types can either be String or Integer.

> **Note**
>
> Custom null values and their data types must match exactly in order for the fields to be recognized as null values and the fields removed. Partial matches where only the custom null value matches but the data type does not will not result in the fields being removed.

Using a SQL query to transform data

You can use a **SQL** transform to write your own transform in the form of a SQL query.

A SQL transform node can have multiple datasets as inputs, but produces only a single dataset as output. In contains a text field, where you enter the Apache SparkSQL query. You can assign aliases to each dataset used as input, to help simply the SQL query. For more information about the SQL syntax, see the Spark SQL documentation.

> **Note**
>
> If you use a Spark SQL transform with a data source located in a VPC, add an AWS Glue VPC endpoint to the VPC that contains the data source. For more information about configuring development endpoints, see Adding a Development Endpoint, Setting Up Your Environment for Development Endpoints, and Accessing Your Development Endpoint in the *AWS Glue Developer Guide*.

To use a SQL transform node in your job diagram

1. (Optional) Add a transform node to the job diagram, if needed. Choose **SQL Query** for the node type.

   > **Note**
   >
   > If you use a data preview session and a custom SQL or custom code node, the data preview session will execute the SQL or code block as-is for the entire dataset.

2. On the **Node properties** tab, enter a name for the node in the job diagram. If a node parent is not already selected, or if you want multiple inputs for the SQL transform, choose a node from the **Node parents** list to use as the input source for the transform. Add additional parent nodes as needed.
3. Choose the **Transform** tab in the node details panel.
4. The source datasets for the SQL query are identified by the names you specified in the **Name** field for each node. If you do not want to use these names, or if the names are not suitable for a SQL query, you can associate a name to each dataset. The console provides default aliases, such as `MyDataSource`.

For example, if a parent node for the SQL transform node is named `Rename Org PK field`, you might associate the name `org_table` with this dataset. This alias can then be used in the SQL query in place of the node name.

5. In the text entry field under the heading **Code block**, paste or enter the SQL query. The text field displays SQL syntax highlighting and keyword suggestions.
6. With the SQL transform node selected, choose the **Output schema** tab, and then choose **Edit**. Provide the columns and data types that describe the output fields of the SQL query.

   Specify the schema using the following actions in the **Output schema** section of the page:

   - To rename a column, place the cursor in the **Key** text box for the column (also referred to as a *field* or *property key*) and enter the new name.
   - To change the data type for a column, select the new data type for the column from the drop-down list.
   - To add a new top-level column to the schema, choose the Overflow (⋯) button, and then choose **Add root key**. New columns are added at the top of the schema.
   - To remove a column from the schema, choose the delete icon (🗑) to the far right of the Key name.
7. When you finish specifying the output schema, choose **Apply** to save your changes and exit the schema editor. If you do not want to save you changes, choose **Cancel** to edit the schema editor.
8. (Optional) After configuring the node properties and transform properties, you can preview the modified dataset by choosing the **Data preview** tab in the node details panel. The first time you choose this tab for any node in your job, you are prompted to provide an IAM role to access the data. There is a cost associated with using this feature, and billing starts as soon as you provide an IAM role.

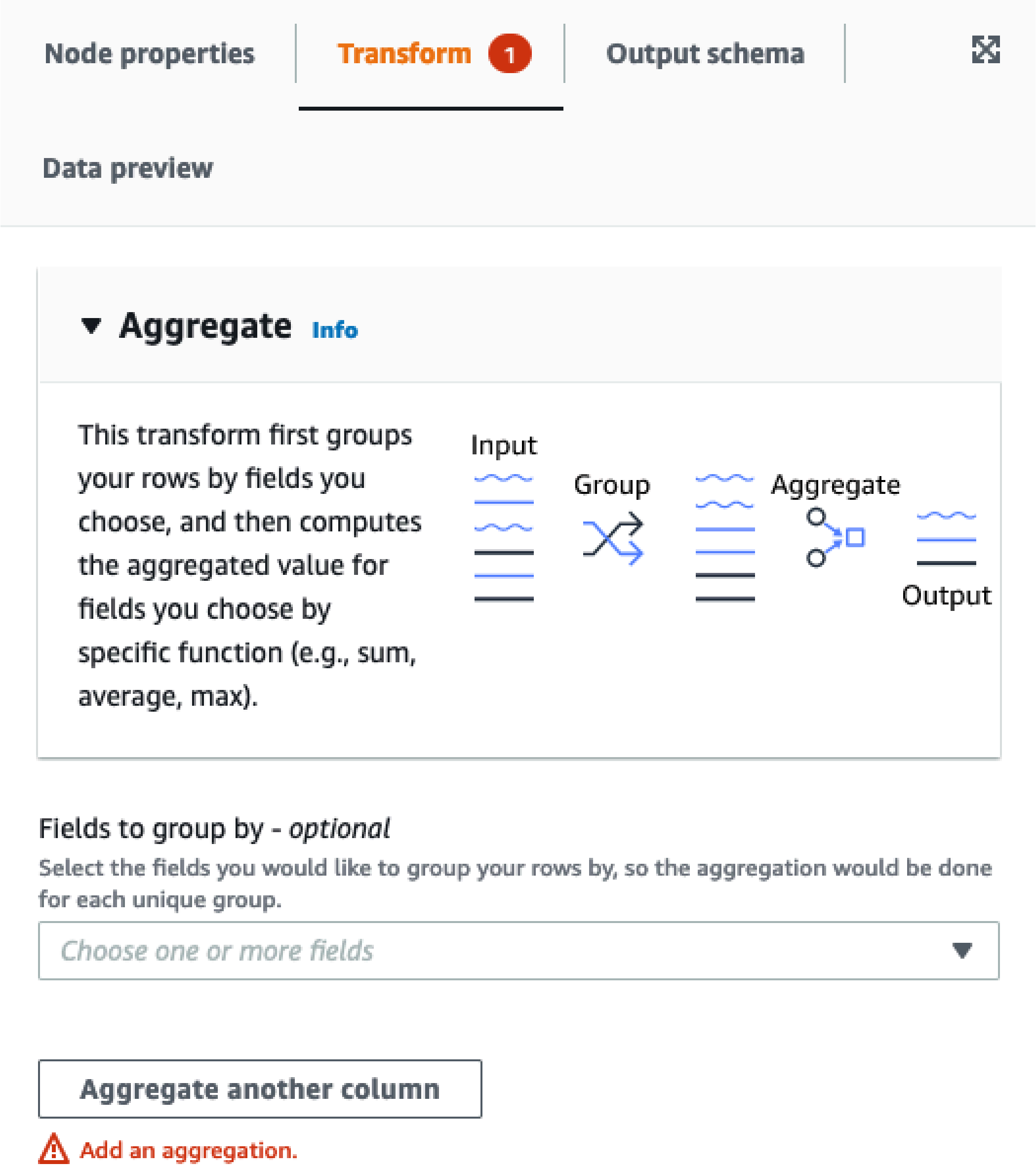
Using Aggregate to perform summary calculations on selected fields

To use the Aggregate transform

1. Add the Aggregate node to the job diagram.

2. On the **Node properties** tab, choose fields to group together by selecting the drop-down field (optional). You can select more than one field at a time or search for a field name by typing in the search bar.

   When fields are selected, the name and datatype are shown. To remove a field, choose 'X' on the field.



Node properties
Transform 1
Output schema
Data preview
▼ Aggregate Info

This transform first groups your rows by fields you choose, and then computes the aggregated value for fields you choose by specific function (e.g., sum, average, max).

Input
Group
Aggregate
Output

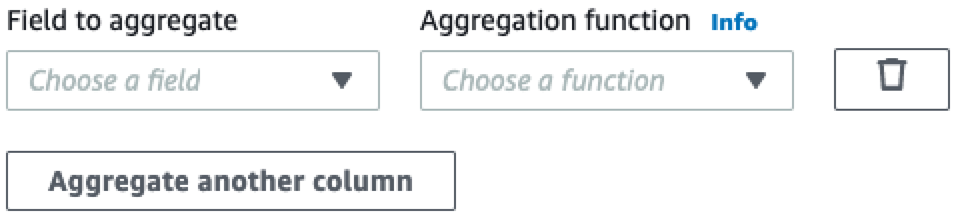
Fields to group by - *optional*

Select the fields you would like to group your rows by, so the aggregation would be done for each unique group.

Choose one or more fields ▼
Aggregate another column

Add an aggregation.

3. Choose **Aggregate another column**. It is required to select at least one field.



4. Choose a field in the **Field to aggregate** drop-down.
5. Choose the aggregation function to apply to the chosen field:

    - avg - calculates the average
    - countDistinct - calculates the number of unique non-null values
    - count - calculates the number of non-null values
    - first - returns the first value that satisfies the 'group by' criteria
    - last - returns the last value that satisfies the 'group by' criteria
    - kurtosis - calculates the the sharpness of the peak of a frequency-distribution curve
    - max - returns the highest value that satisfies the 'group by' criteria
    - min - returns the lowest value that satisfies the 'group by' criteria
    - skewness - measure of the asymmetry of the probability distribution of a normal distribution
    - stddev_pop - calculates the population standard deviation and returns the square root of the population variance
    - sum - the sum of all values in the group
    - sumDistinct - the sum of distinct values in the group
    - var_samp - the sample variance of the group (ignores nulls)
    - var_pop - the population variance of the group (ignores nulls)

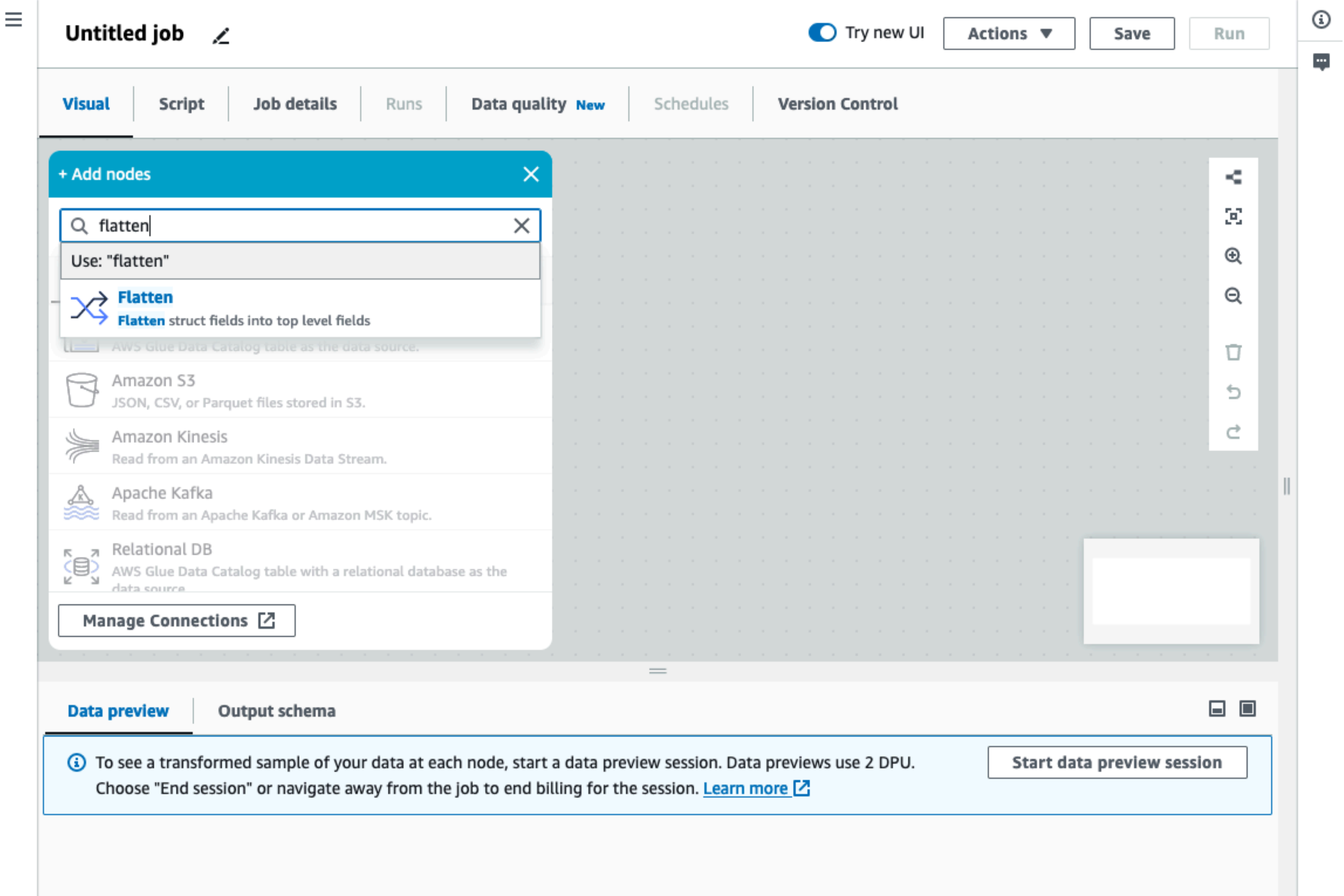
Flatten nested structs

Flatten the fields of nested structs in the data, so they become top level fields. The new fields are named using the field name prefixed with the names of the struct fields to reach it, separated by dots.

For example, if the data has a field of type Struct named “phone_numbers”, which among other fields has one of type “Struct” named “home_phone” with two fields: “country_code” and “number”. Once flattened, these two fields will become top level fields named: “phone_numbers.home_phone.country_code” and “phone_numbers.home_phone.number” respectively.

To add a *Flatten* transform node in your job diagram

1. Open the Resource panel and then choose the **Transforms** tab, then **Flatten** to add a new transform to your job diagram. You can also use the search bar by entering 'Flatten', then clicking the Flatten node. The node selected at the time of adding the node will be its parent.


Untitled job
Try new UI
Actions
Save
Run
Visual
Script
Job details
Runs
Data quality New
Schedules
Version Control
+ Add nodes
flatten
Use: "flatten"
Flatten
Flatten struct fields into top level fields
Amazon S3
JSON, CSV, or Parquet files stored in S3.
Amazon Kinesis
Read from an Amazon Kinesis Data Stream.
Apache Kafka
Read from an Apache Kafka or Amazon MSK topic.
Relational DB
AWS Glue Data Catalog table with a relational database as the data source
Manage Connections
Data preview
Output schema
Start data preview session

To see a transformed sample of your data at each node, start a data preview session. Data previews use 2 DPU. Choose "End session" or navigate away from the job to end billing for the session. Learn more

2. (Optional) On the **Node properties** tab, you can enter a name for the node in the job diagram. If a node parent is not already selected, then choose a node from the **Node parents** list to use as the input source for the transform.

3. (Optional) On the **Transform** tab, you can limit the maximum nesting level to flatten. For instance, setting that value to 1 means that only top-level structs will be flattened. Setting the max to 2 will flatten the top level and the structs directly under it.

Add a UUID column

When you add a *UUID* (Universally Unique Identified) column, each row will be assigned a unique 36-character string.

To add a *UUID* transform node in your job diagram

1. Open the Resource panel and then choose **UUID** to add a new transform to your job diagram. The node selected at the time of adding the node will be its parent.
2. (Optional) On the **Node properties** tab, you can enter a name for the node in the job diagram. If a node parent is not already selected, then choose a node from the **Node parents** list to use as the input source for the transform.
3. (Optional) On the **Transform** tab, you can customize the name of the new column. By default it will be named "uuid".

Add an identifier column

Assign a numeric *Identifier* for each row in the dataset.

To add an *Identifier* transform node in your job diagram

1. Open the Resource panel and then choose **Identifier** to add a new transform to your job diagram. The node selected at the time of adding the node will be its parent.
2. (Optional) On the **Node properties** tab, you can enter a name for the node in the job diagram. If a node parent is not already selected, then choose a node from the **Node parents** list to use as the input source for the transform.
3. (Optional) On the **Transform** tab, you can customize the name of the new column. By default, it will be named "id".
4. (Optional) If the job processes and stores data incrementally, you want to avoid the same ids to be reused between job runs.

   On the **Transform** tab, mark the **unique** checkbox option. It will include the job timestamp in the identifier, making it unique between multiple runs. To allow for the larger number, the column instead of type long will be a decimal.

Convert a column to timestamp type

You can use the transform *To timestamp* to change the data type of a numeric or string column into timestamp, so that it can be stored with that data type or applied to other transforms that require a timestamp.

To add a *To timestamp* transform node in your job diagram

1. Open the Resource panel and then choose **To timestamp** to add a new transform to your job diagram. The node selected at the time of adding the node will be its parent.
2. (Optional) On the **Node properties** tab, you can enter a name for the node in the job diagram. If a node parent is not already selected, then choose a node from the **Node parents** list to use as the input source for the transform.
3. On the **Transform** tab, enter the name of the column to be converted.
4. On the **Transform** tab, define how to parse the column selected by choosing the type.

   If the value is a number, it can be expressed in seconds (Unix/Python timestamp), milliseconds or microseconds, choose the corresponding option.

   If the value is a formatted string, choose the "iso" type, the string needs to conform to one of the variants of the ISO format, for example: “2022-11-02T14:40:59.915Z“.

   If you don’t know the type at this point or different rows use different types, then you can choose ”autodetect“ and the system will make its best guess, with a small performance cost.
5. (Optional) On the **Transform** tab, instead of converting the selected column, you can create a new one and keep the original by entering a name for the new column.

Convert a timestamp column to a formatted string

Format a timestamp column into a string based on a pattern. You can use *Format timestamp* to get date and time as a string with the desired format. You can define the format using Spark date syntax as well as most of the Python date codes.

For example, if you want your date string to be formatted like “2023-01-01 00:00”, you can define such format using the Spark syntax as “yyyy-MM-dd HH:mm” or the equivalent Python date codes as “%Y-%m-%d %H:%M”

To add a *Format timestamp* transform node in your job diagram

1. Open the Resource panel and then choose **Format timestamp** to add a new transform to your job diagram. The node selected at the time of adding the node will be its parent.
2. (Optional) On the **Node properties** tab, you can enter a name for the node in the job diagram. If a node parent is not already selected, then choose a node from the **Node parents** list to use as the input source for the transform.
3. On the **Transform** tab, enter the name of the column to be converted.
4. On the **Transform** tab, enter the **Timestamp format** pattern to use, expressed using Spark date syntax or Python date codes.
5. (Optional) On the **Transform** tab, instead of converting the selected column, you can create a new one and keep the original by entering a name for the new column.

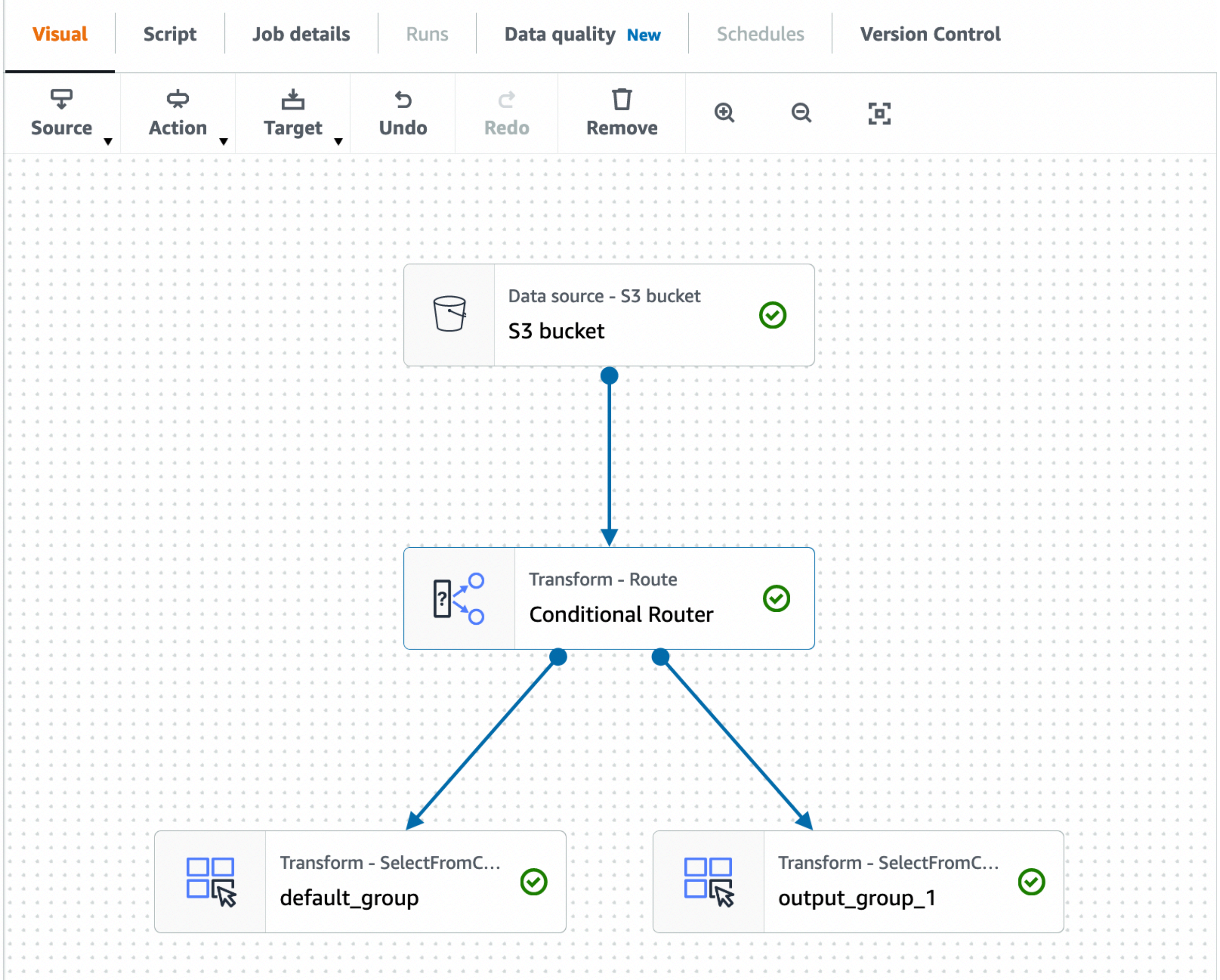
Creating a Conditional Router transformation

The Conditional Router transform allows you to apply multiple conditions to incoming data. Each row of the incoming data is evaluated by a group filter condition and processed into its corresponding group. If a row meets more than one group filter condition, the transform passes the row to multiple groups. If a row does not meet any condition, it can either be dropped or routed to a default output group.

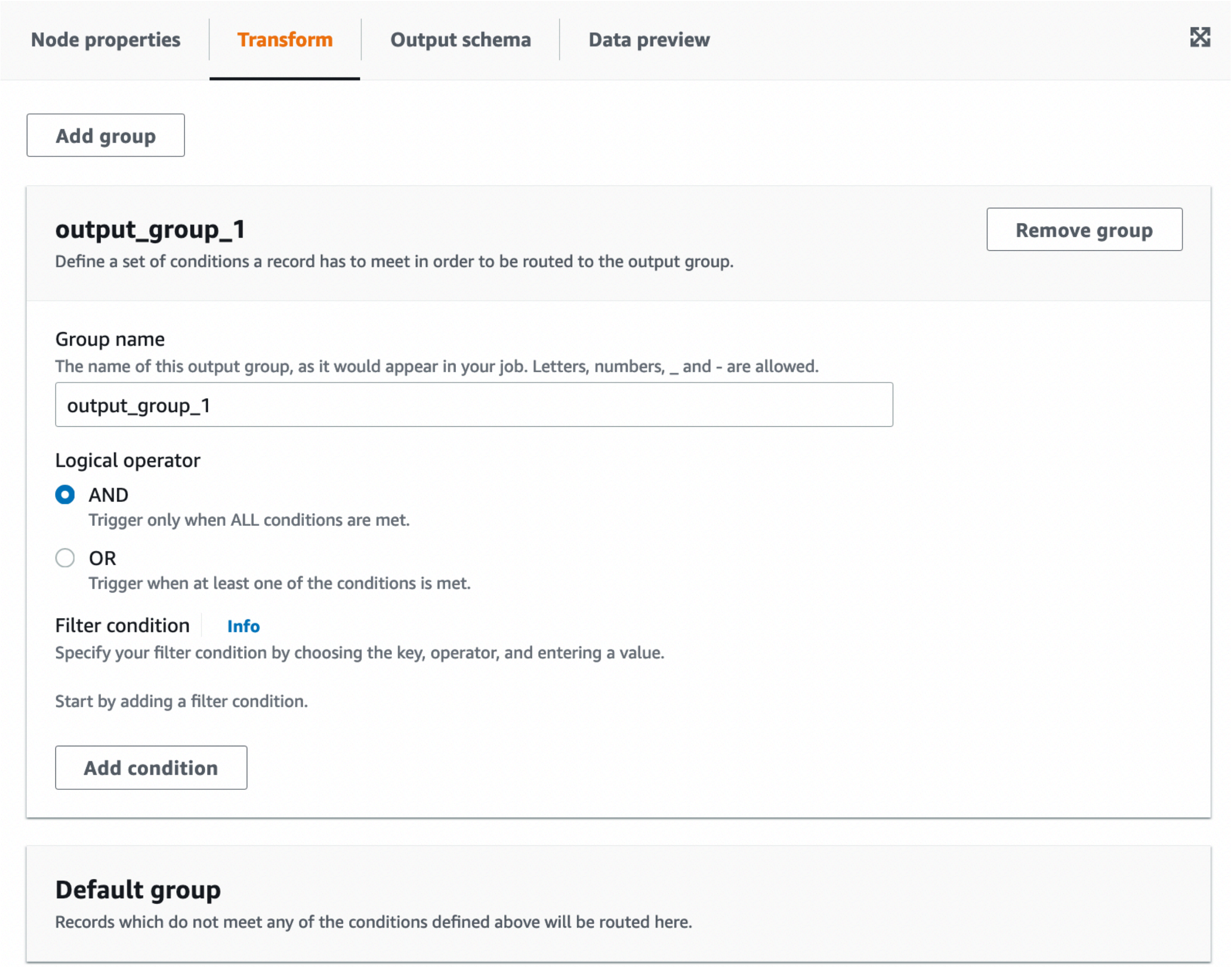
This transform is similar to the filter transform, but useful for users who want to test the same input data on multiple conditions.

To add a Conditional Router transform:

1. Choose a node where you will perform the conditional router transformation. This can be a source node or another transform.
2. Choose **Action**, then use the search bar to find and choose 'Conditional Router'. A **Conditional Router** transform is added along with two output nodes. One output node, 'Default group', contains records which do not meet any of the conditions defined in the other output node(s). The default group cannot be edited.


Visual
Script
Job details
Runs
Data quality New
Schedules
Version Control
Source
Action
Target
Undo
Redo
Remove
Data source - S3 bucket
S3 bucket
Transform - Route
Conditional Router
Transform - SelectFromC...
default_group
Transform - SelectFromC...
output_group_1

You can add additional output groups by choosing **Add group**. For each output group, you can name the group and add filter conditions and a logical operator.



output_group_1

Define a set of conditions a record has to meet in order to be routed to the output group.

Group name

The name of this output group, as it would appear in your job. Letters, numbers, _ and - are allowed.

output_group_1

Logical operator

● AND

Trigger only when ALL conditions are met.

○ OR

Trigger when at least one of the conditions is met.

Filter condition

Specify your filter condition by choosing the key, operator, and entering a value.

Start by adding a filter condition.

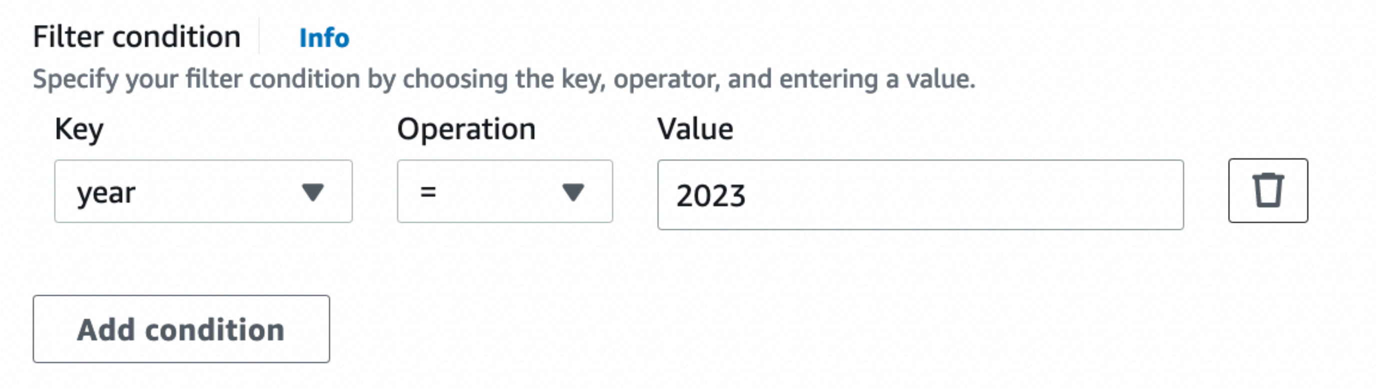
Default group

Records which do not meet any of the conditions defined above will be routed here.

3. Rename the output group name by entering a new name for the group. AWS Glue Studio will automatically name your groups for you (for example, 'output_group_1').

4. Choose a logical operator (**AND**, **OR**) and add a **Filter condition** by specifying the **Key**, **Operation**, and **Value**. Logical operators allow you to implement more than one filter condition and perform the logical operator on each filter condition you specify.

   When specifying the key, you can choose from available keys in your schema. You can then choose the available operation depending on the type of key you selected. For example, if the key type is 'string', then the available operation to choose from is 'matches'.


Filter condition Info

Specify your filter condition by choosing the key, operator, and entering a value.

Key: year

Operation: =

Value: 2023

5. Enter the value in the **Value** field. To add additional filter conditions, choose **Add condition**. To remove filter conditions, choose the trash can icon.

Using the Concatenate Columns transform to append columns

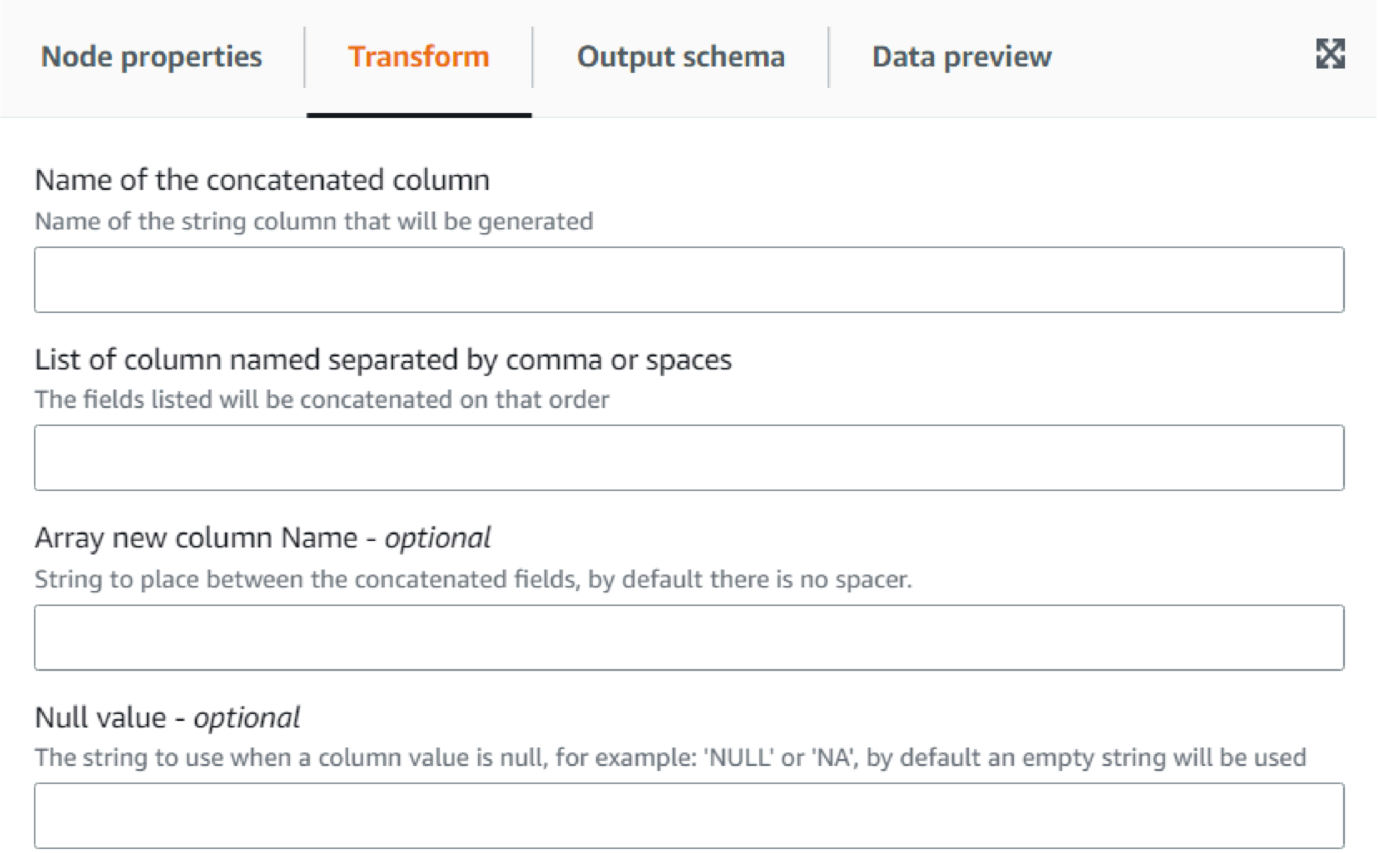
The Concatenate transform allows you to build a new string column using the values of other columns with an optional spacer. For example, if we define a concatenated column “date” as the concatenation of “year”, “month” and “day” (in that order) with “-” as the spacer, we would get:

day	month	year	date
01	01	2020	2020-01-01
02	01	2020	2020-01-02
03	01	2020	2020-01-03
04	01	2020	2020-01-04

To add a Concatenate transform:

1. Open the Resource panel. Then choose **Concatenate Columns** to add a new transform to your job diagram. The node selected at the time of adding the node will be its parent.
2. (Optional) On the **Node properties** tab, you can enter a name for the node in the job diagram. If a node parent is not already selected, then choose a node from the Node parents list to use as the input source for the transform.

3. On the **Transform** tab, enter the name of the column that will hold the concatenated string as well as the columns to concatenate. The order in which you check the columns in the dropdown will be the order used.



Name of the concatenated column

Name of the string column that will be generated

List of column named separated by comma or spaces

The fields listed will be concatenated on that order

Array new column Name - *optional*

String to place between the concatenated fields, by default there is no spacer.

Null value - *optional*

The string to use when a column value is null, for example: 'NULL' or 'NA', by default an empty string will be used

4. **Spacer - optional** – Enter a string to place betwen the concatenated fields. By default, there is no spacer.

5. **Null value - optional** – Enter a string to use when a column value is null. By default, in the cases where columns have the value 'NULL' or 'NA', an empty string is used.

Using the Split String transform to break up a string column

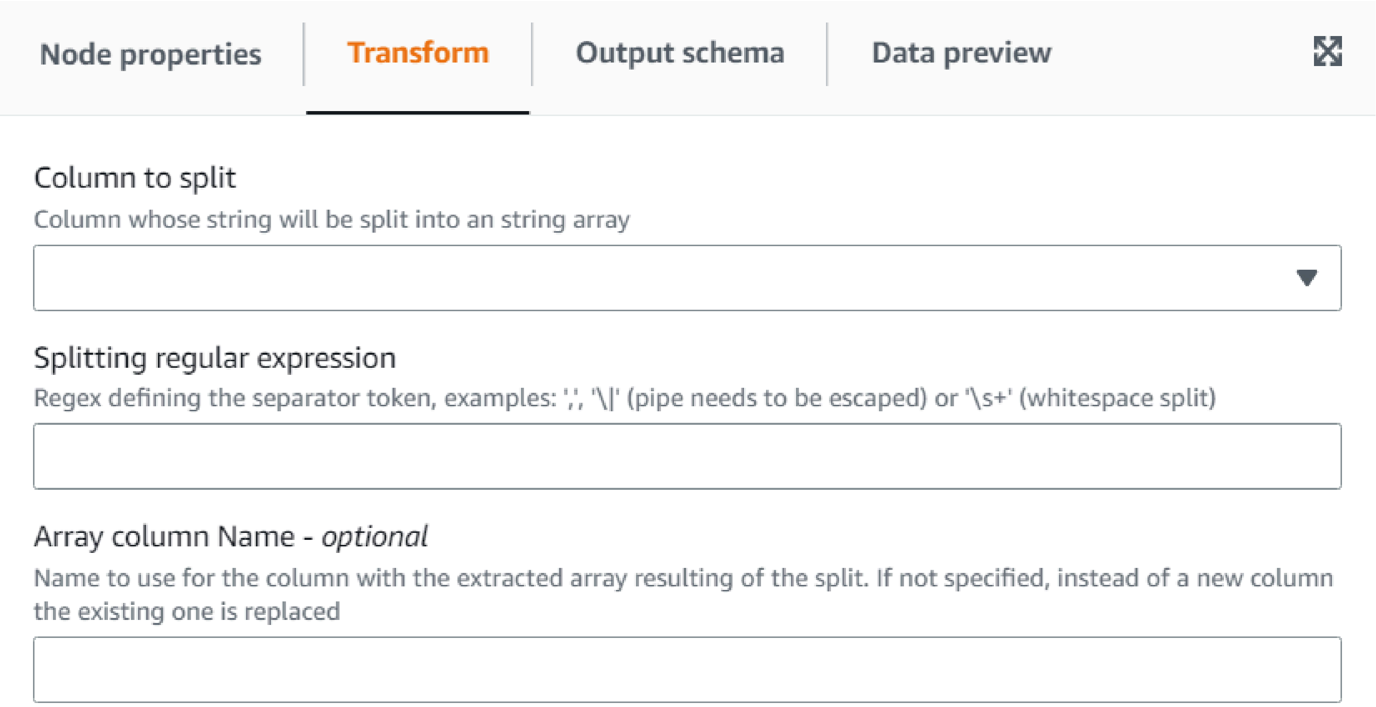
The Split String transform allows you to break up a string into an array of tokens using a regular expression to define how the split is done. You can then keep the column as an array type or apply an **Array To Columns** transform after this one, to extract the array values onto top level fields, assuming that each token has a meaning we know beforehand. Also, if the order of the tokens is irrelevant (for instance, a set of categories), you can use the **Explode** transform to generate a separate row for each value.

For example, you can split a the column “categories” using a comma as a pattern to add a column “categories_arr”.

product_id	categories	categories_arr
1	sports,winter	[sports, winter]
2	garden,tools	[garden, tools]
3	videogames	[videogames]
4	game,boardgame,social	[game, boardgame, social]

To add a Split String transform:

1. Open the Resource panel and then choose Split String to add a new transform to your job diagram. The node selected at the time of adding the node will be its parent.
2. (Optional) On the Node properties tab, you can enter a name for the node in the job diagram. If a node parent is not already selected, then choose a node from the Node parents list to use as the input source for the transform.
3. On the **Transform** tab, choose the column to split and enter the pattern to use to split the string. In most cases you can just enter the character(s) unless it has a special meaning as a regular expression and needs to be escaped. The characters that need escaping are: `\.[]{}()<>*+-=!?^$|` by adding a backslash in front of the character. For instance if you want to separate by a dot ('.') you need to enter `\.`. However, a comma doesn’t have a special meaning and can just be specified as is: `,`.



Column to split

Column whose string will be split into an string array

Splitting regular expression

Regex defining the separator token, examples: ',', '\|' (pipe needs to be escaped) or '\s+' (whitespace split)

Array column Name - *optional*

Name to use for the column with the extracted array resulting of the split. If not specified, instead of a new column the existing one is replaced

4. (Optional) If you want to keep the original string column, then you can enter a name for a new array column, this way keeping both the original string column and the new tokenized array column.

Using the Array To Columns transform to extract the elements of an array into top level columns

The Array To Columns transform allows you extract some or all the elements of a column of type array into new columns. The transform will fill the new columns as much as possible if the array has enough values to extract, optionally taking the elements in the positions specified.

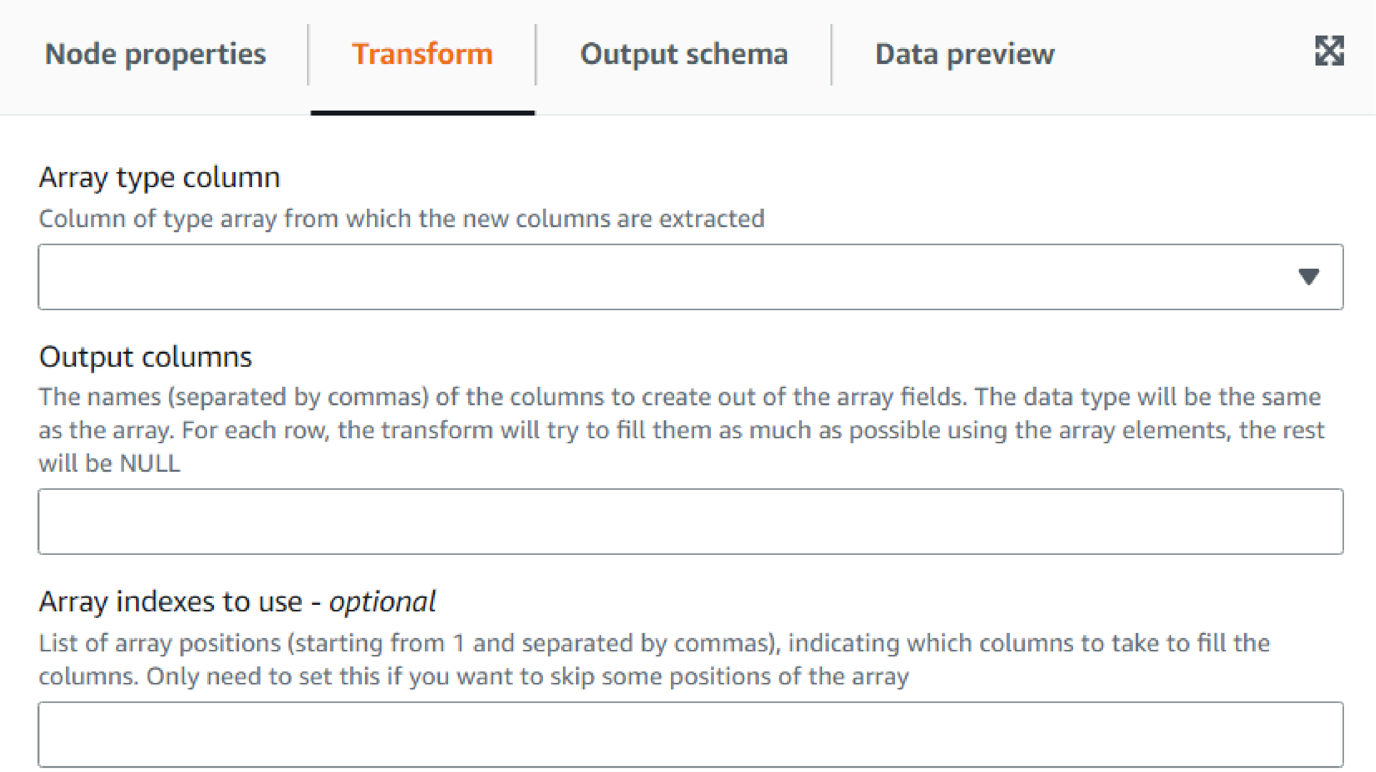
For instance, if you have an array column “subnet”, which was the result of applying the “Split String” transform on a ip v4 subnet, you can extract the first and forth positions into new columns “first_octect” and “forth_octect”. The output of the transform in this example would be (notice the last two rows have shorter arrays than expected):

subnet	first_octect	fourth_octect
[54, 240, 197, 238]	54	238
[192, 168, 0, 1]	192	1

subnet	first_octect	fourth_octect
[192, 168]	192	
[]		

To add a Array To Columns transform:

1. Open the Resource panel and then choose **Array To Columns** to add a new transform to your job diagram. The node selected at the time of adding the node will be its parent.
2. (Optional) On the **Node properties** tab, you can enter a name for the node in the job diagram. If a node parent is not already selected, then choose a node from the Node parents list to use as the input source for the transform.
3. On the **Transform** tab, choose the array column to extract and enter the list of new columns for the tokens extracted.



Array type column

Column of type array from which the new columns are extracted

Output columns

The names (separated by commas) of the columns to create out of the array fields. The data type will be the same as the array. For each row, the transform will try to fill them as much as possible using the array elements, the rest will be NULL

Array indexes to use - *optional*

List of array positions (starting from 1 and separated by commas), indicating which columns to take to fill the columns. Only need to set this if you want to skip some positions of the array

4. (Optional) If you don’t want to take the array tokens in order to assign to columns, you can specify the indexes to take which will be assigned to the list of columns in the same order specified. For instance if the output columns are “column1, column2, column3” and the

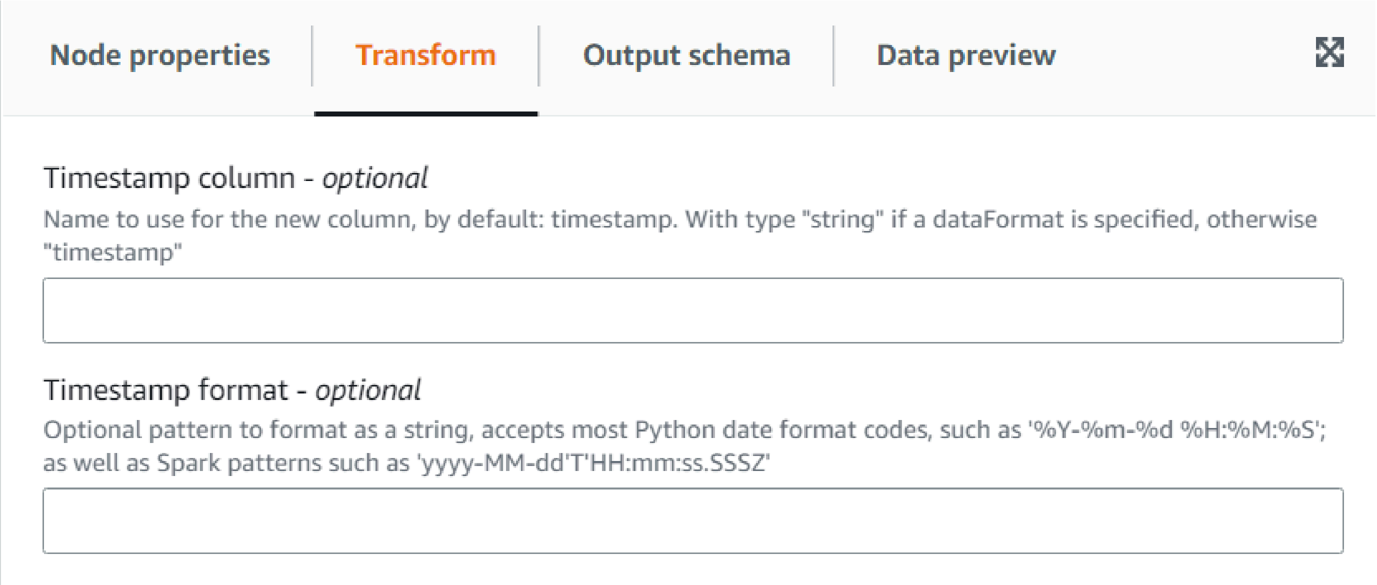
indexes “4, 1, 3”, the forth element of the array will go to column1, the first to column2 and the third to column3 (if the array is shorter than the index number, a NULL value will be set).

Using the Add Current Timestamp transform

The **Add Current Timestamp** transform allows you to mark the rows with the time on which the data was processed. This is useful for auditing purposes or to track latency in the data pipeline. You can add this new column as a timestamp data type or a formatted string.

To add a Add Current Timestamp transform:

1. Open the Resource panel and then choose **Add Current Timestamp** to add a new transform to your job diagram. The node selected at the time of adding the node will be its parent.
2. (Optional) On the **Node properties** tab, you can enter a name for the node in the job diagram. If a node parent is not already selected, then choose a node from the Node parents list to use as the input source for the transform.



Timestamp column - *optional*

Name to use for the new column, by default: timestamp. With type "string" if a dataFormat is specified, otherwise "timestamp"

Timestamp format - *optional*

Optional pattern to format as a string, accepts most Python date format codes, such as '%Y-%m-%d %H:%M:%S'; as well as Spark patterns such as 'yyyy-MM-dd'T'HH:mm:ss.SSSZ'

3. (Optional) On the **Transform** tab, enter a custom name for the new column and a format if you rather the column to be a formatted date string.

Using the Pivot Rows to Columns transform

The **Pivot Rows to Columns** transform allows you to aggregate a numeric column by rotating unique values on selected columns which become new columns (if multiple columns are selected, the values are concatenated to name the new columns). That way rows are consolidated while having more columns with partial aggregations for each unique value. For example, if you have this dataset of sales by month and country (sorted to be easier to illustrate):

year	month	country	amount
2020	Jan	uk	32
2020	Jan	de	42
2020	Jan	us	64
2020	Feb	uk	67
2020	Feb	de	4
2020	Feb	de	7
2020	Feb	us	6
2020	Feb	us	12
2020	Jan	us	90

If you pivot **amount** and **country** as the aggregation columns, new columns are created from the original **country** column. In the table below, you have new columns for **de**, **uk**, and **us** instead of the **country** column.

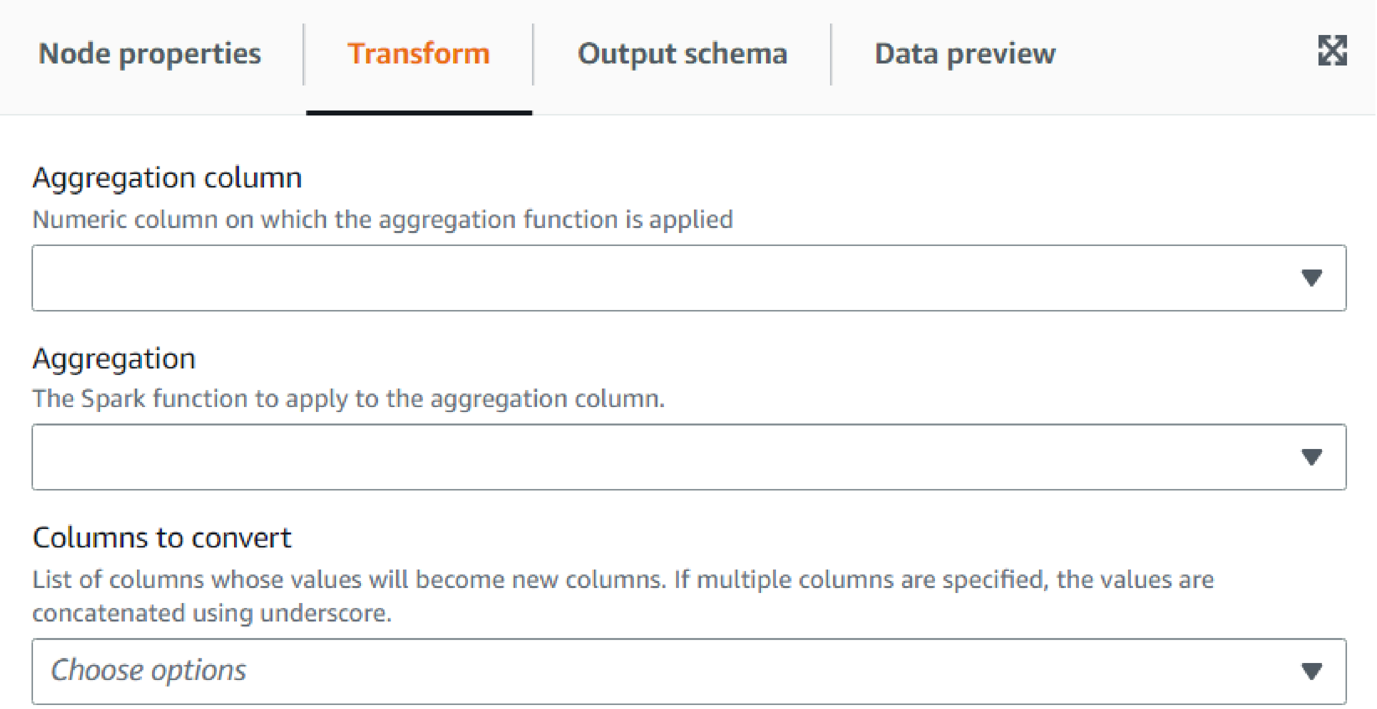
year	month	de	uk	us
2020	Jan	42	32	64
2020	Jan	11	67	18
2021	Jan			90

If instead you want to pivot both the month and county, you get a column for each combination of the values of those columns:

year	Jan_de	Jan_uk	Jan_us	Feb_de	Feb_uk	Feb_us
2020	42	32	64	11	67	18
2021			90			

To add a Pivot Rows To Columns transform:

1. Open the Resource panel and then choose **Pivot Rows To Columns** to add a new transform to your job diagram. The node selected at the time of adding the node will be its parent.
2. (Optional) On the **Node properties** tab, you can enter a name for the node in the job diagram. If a node parent is not already selected, then choose a node from the Node parents list to use as the input source for the transform.
3. On the **Transform** tab, choose the numeric column which will be aggregated to produce the values for the new columns, the aggregation function to apply and the column(s) to convert its unique values into new columns.



Aggregation column

Numeric column on which the aggregation function is applied

Aggregation

The Spark function to apply to the aggregation column.

Columns to convert

List of columns whose values will become new columns. If multiple columns are specified, the values are concatenated using underscore.

Choose options

Using the Unpivot Columns To Rows transform

The **Unpivot** transform allows you convert columns into values of new columns generating a row for each unique value. It’s the opposite of pivot but note that it’s not equivalent since it cannot separate rows with identical values that were aggregated or split combinations into the original columns (you can do that later using a Split transform). For example, if you have the following table:

year	month	de	uk	us
2020	Jan	42	32	64
2020	Feb	11	67	18
2021	Jan			90

You can unpivot the columns: “de”, “uk” and “us” into a column “country” with the value “amount”, and get the following (sorted here for illustration purposes):

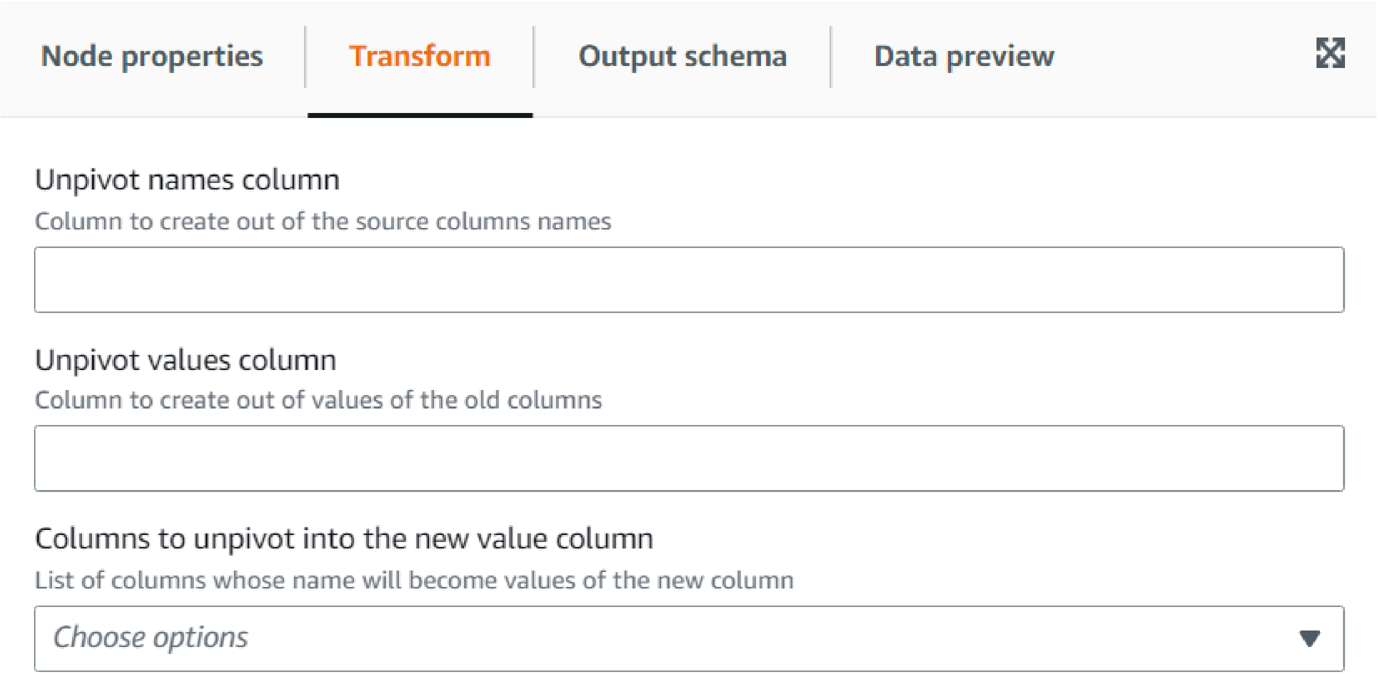
year	month	country	amount
2020	Jan	uk	32
2020	Jan	de	42
2020	Jan	us	64
2020	Feb	uk	67
2020	Feb	de	11
2020	Feb	us	18
2021	Jan	us	90

Notice the columns that have a NULL value (“de” and “uk of Jan 2021) don’t get generated by default. You can enable that option to get:

year	month	country	amount
2020	Jan	uk	32
2020	Jan	de	42
2020	Jan	us	64
2020	Feb	uk	67
2020	Feb	de	11
2020	Feb	us	18
2021	Jan	us	90
2021	Jan	de	
2021	Jan	uk	

To add a Unpivot Columns to Rows transform:

1. Open the Resource panel and then choose **Unpivot Columns to Rows** to add a new transform to your job diagram. The node selected at the time of adding the node will be its parent.
2. (Optional) On the **Node properties** tab, you can enter a name for the node in the job diagram. If a node parent is not already selected, then choose a node from the Node parents list to use as the input source for the transform.
3. On the **Transform** tab, enter the new columns to be created to hold the names and values of the columns chosen to unpivot.



Unpivot names column

Column to create out of the source columns names

Unpivot values column

Column to create out of values of the old columns

Columns to unpivot into the new value column

List of columns whose name will become values of the new column

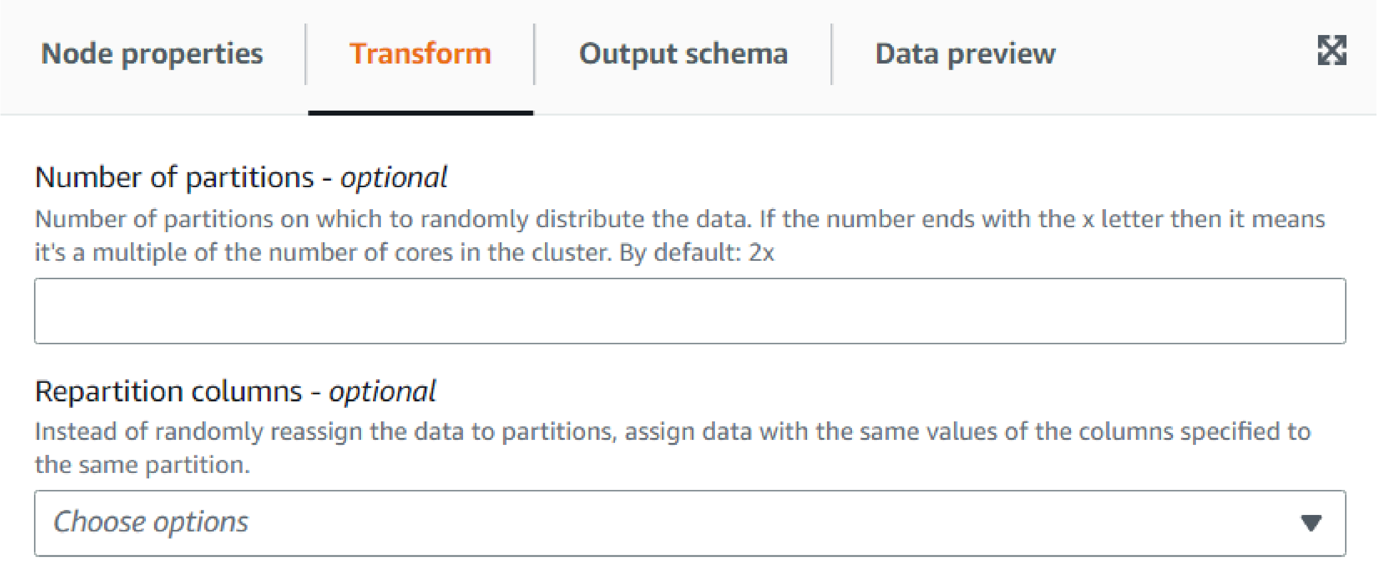
Choose options

Using the Autobalance Processing transform to optimize your runtime

The **Autobalance Processing** transform redistributes the data among the workers for better performance. This helps in cases where the data is unbalanced or as it comes from the source doesn’t allow enough parallel processing on it. This is common where the source is gzipped or is JDBC. The redistribution of data has a modest performance cost, so the optimization might not always compensate that effort if the data was already well balanced. Underneath, the transform uses Apache Spark repartition to randomly reassign data among a number of partitions optimal for the cluster capacity. For advanced users, it’s possible to enter a number of partitions manually. In addition, it can be used to optimize the writing of partitioned tables by reorganizing the data based on specified columns. This results in output files that are more consolidated.

1. Open the Resource panel and then choose **Autobalance Processing** to add a new transform to your job diagram. The node selected at the time of adding the node will be its parent.
2. (Optional) On the **Node properties** tab, you can enter a name for the node in the job diagram. If a node parent is not already selected, then choose a node from the Node parents list to use as the input source for the transform.
3. (Optional) On the **Transform** tab, you can enter a number of partitions. In general, it’s recommended that you let the system decide this value, however you can tune the multiplier or enter a specific value if you need to control this. If you are going to save the data

partitioned by columns, you can choose the same columns as repartition columns. This way it will minimize the number of files on each partition and avoid having many files per partitions, which would hinder the performance of the tools querying that data.



Number of partitions - *optional*

Number of partitions on which to randomly distribute the data. If the number ends with the x letter then it means it's a multiple of the number of cores in the cluster. By default: 2x

Repartition columns - *optional*

Instead of randomly reassign the data to partitions, assign data with the same values of the columns specified to the same partition.

Using the Derived Column transform to combine other columns

The **Derived Column** transform allows you to define a new column based on a math formula or SQL expression in which you can use other columns in the data, as well as constants and literals. For instance, to derive a “percentage” column from the columns "success" and "count", you can enter the SQL expression: "success * 100 / count || '%'".

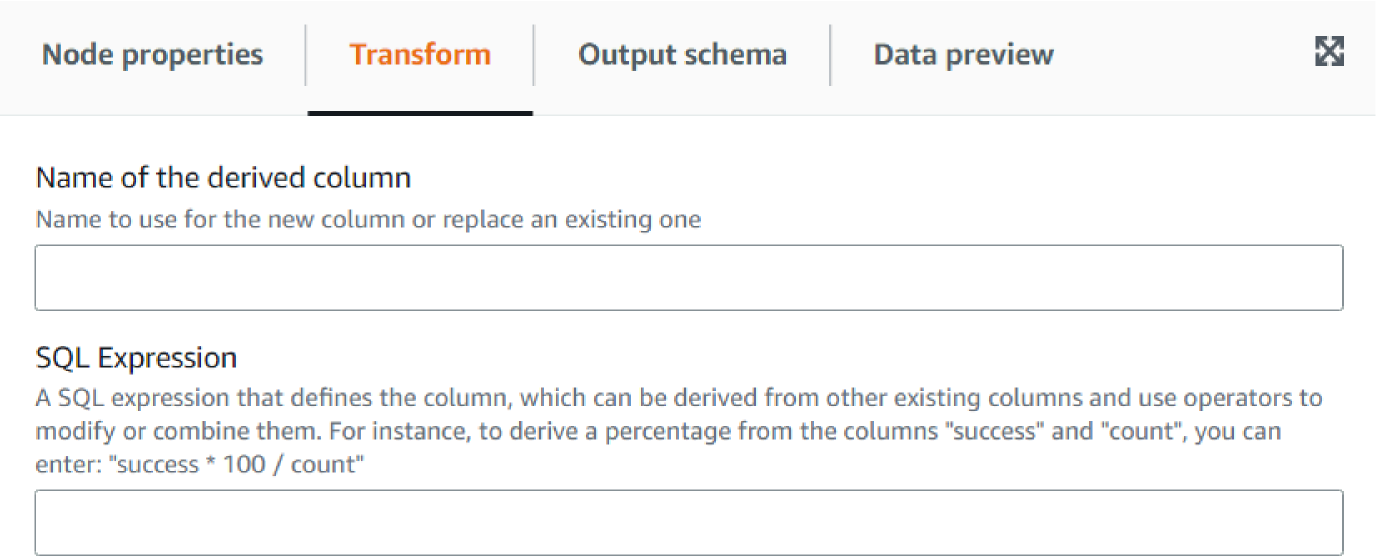
Example result:

success	count	percentage
14	100	14%
6	20	3%
3	40	7.5%

To add a Derived Column transform:

1. Open the Resource panel and then choose **Derived Column** to add a new transform to your job diagram. The node selected at the time of adding the node will be its parent.

2. (Optional) On the **Node properties** tab, you can enter a name for the node in the job diagram. If a node parent is not already selected, then choose a node from the Node parents list to use as the input source for the transform.
3. On the **Transform** tab, enter the name of the column and the expression for its content.



Name of the derived column

Name to use for the new column or replace an existing one

SQL Expression

A SQL expression that defines the column, which can be derived from other existing columns and use operators to modify or combine them. For instance, to derive a percentage from the columns "success" and "count", you can enter: "success * 100 / count"

Using the Lookup transform to add matching data from a catalog table

The **Lookup** transform allows you to add columns from a defined catalog table when the keys match the defined lookup columns in the data. This is equivalent to doing a left outer join between the data and the lookup table using as condition matching columns.

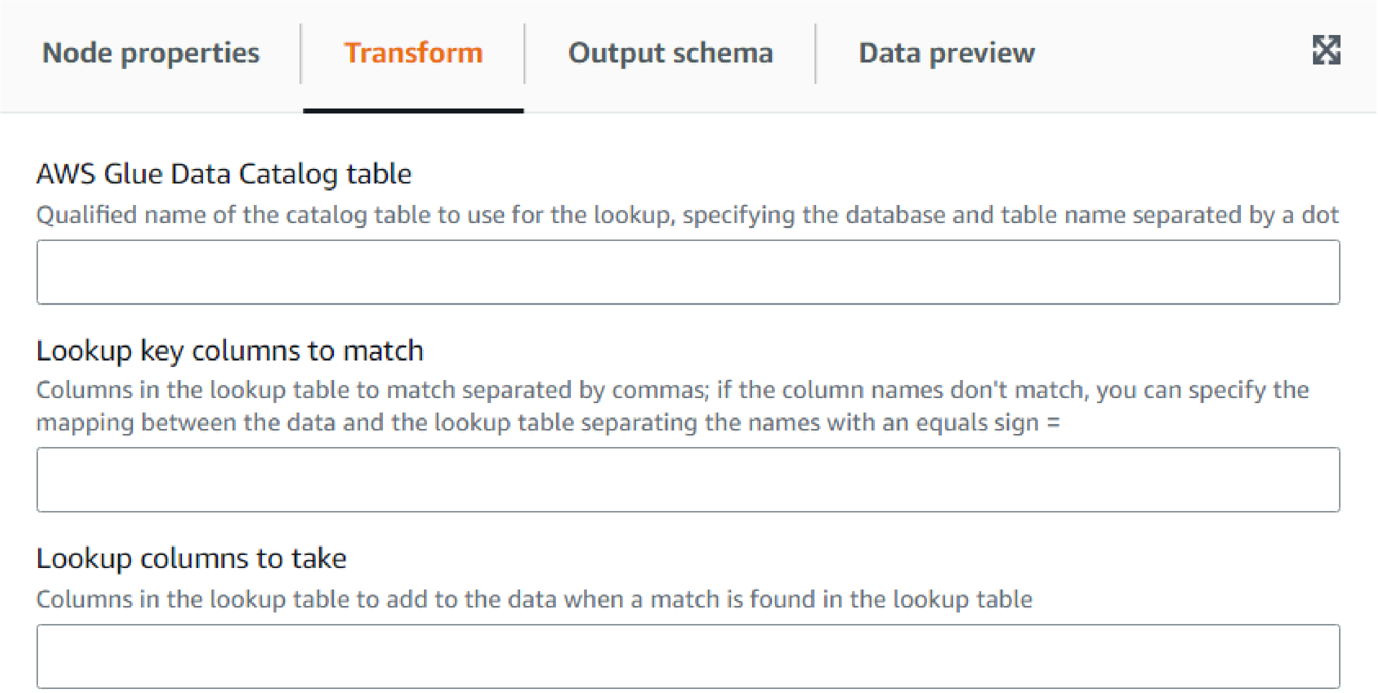
To add a Lookup transform:

1. Open the Resource panel and then choose **Lookup** to add a new transform to your job diagram. The node selected at the time of adding the node will be its parent.
2. (Optional) On the **Node properties** tab, you can enter a name for the node in the job diagram. If a node parent is not already selected, then choose a node from the Node parents list to use as the input source for the transform.
3. On the **Transform** tab, enter the fully qualified catalog table name to use to perform the lookups. For example, if your database is “mydb” and your table “mytable” then enter “mydb.mytable”. Then enter the criteria to find a match in the lookup table, if the lookup key is composed. Enter the list of key columns separated by commas. If one or more of the key columns don’t have the same name then you need to define the match mapping.

For example, if the data columns are “user_id” and “region” and in the users table the corresponding columns are named “id” and “region“, then in the **Columns to match** field, enter: ”user_id=id, region“. You could do region=region but it’s not needed since they are the same.

4. Finally, enter the columns to bring from the row matched in the lookup table to incorporate them into the data. If no match was found those columns will be set to NULL.

> **Note**
>
> Underneath the **Lookup** transform, it is using a left join in order to be efficient. If the lookup table has a composite key, ensure the columns to match are setup to match all the key columns so that only one match can occur. Otherwise, multiple lookup rows will match and this will result in extra rows added for each of those matches.



AWS Glue Data Catalog table

Qualified name of the catalog table to use for the lookup, specifying the database and table name separated by a dot

Lookup key columns to match

Columns in the lookup table to match separated by commas; if the column names don't match, you can specify the mapping between the data and the lookup table separating the names with an equals sign =

Lookup columns to take

Columns in the lookup table to add to the data when a match is found in the lookup table

Using the Explode Array or Map Into Rows transform

The **Explode** transform allows you to extract values from a nested structure into individual rows that are easier to manipulate. In the case of an array, the transform will generate a row for each value of the array, replicating the values for the other columns in the row. In the case of a map, the transform will generate a row for each entry with the key and value as columns plus any other columns in the row.

For example, if we have this dataset which has a “category” array column with multiple values.

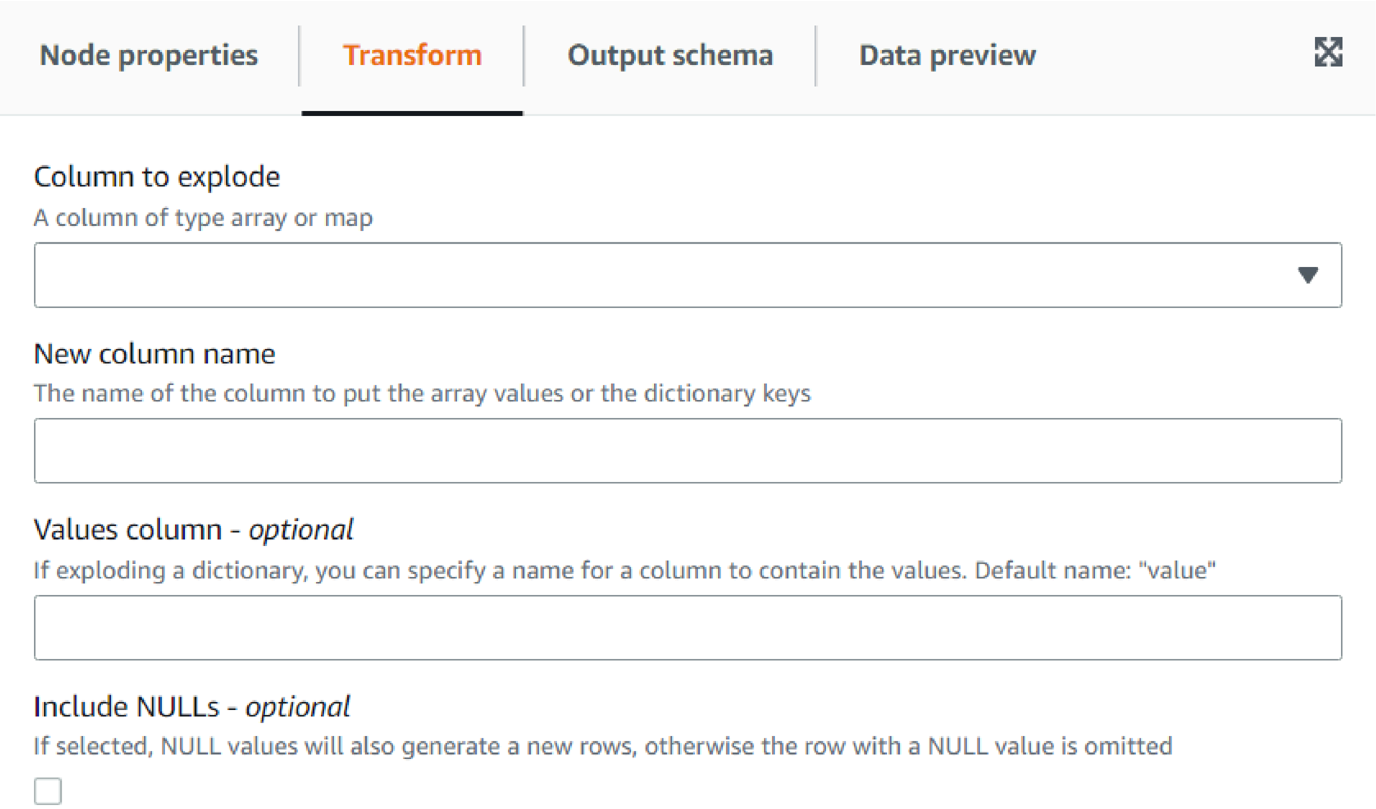
product_id	category
1	[sports, winter]
2	[garden, tools]
3	[videogames]
4	[game, boardgame, social]
5	[]

If you explode the 'category' column into a column with the same name, you will override the column. You can select that you want NULLs included to get the following (ordered for illustration purposes):

product_id	category
1	sports
1	winter
2	garden
2	tool
3	videogames
4	game
4	boardgame
4	social
5	

To add a Explode Array Or Map Into Rows transform:

1. Open the Resource panel and then choose **Explode Array Or Map Into Rows** to add a new transform to your job diagram. The node selected at the time of adding the node will be its parent.
2. (Optional) On the **Node properties** tab, you can enter a name for the node in the job diagram. If a node parent is not already selected, then choose a node from the Node parents list to use as the input source for the transform.
3. On the **Transform** tab, choose the column to explode (it must be an array or map type). Then enter a name for the column for the items of the array or the names of the columns for the keys and values if you are exploding a map.
4. (Optional) On the **Transform** tab, by default if the column to explode is NULL or has an empty structure, it will be omitted on the exploded dataset. If you want to keep the row (with the new columns as NULL) then check “Include NULLs”.



Column to explode

A column of type array or map

New column name

The name of the column to put the array values or the dictionary keys

Values column - *optional*

If exploding a dictionary, you can specify a name for a column to contain the values. Default name: "value"

Include NULLs - *optional*

If selected, NULL values will also generate a new rows, otherwise the row with a NULL value is omitted

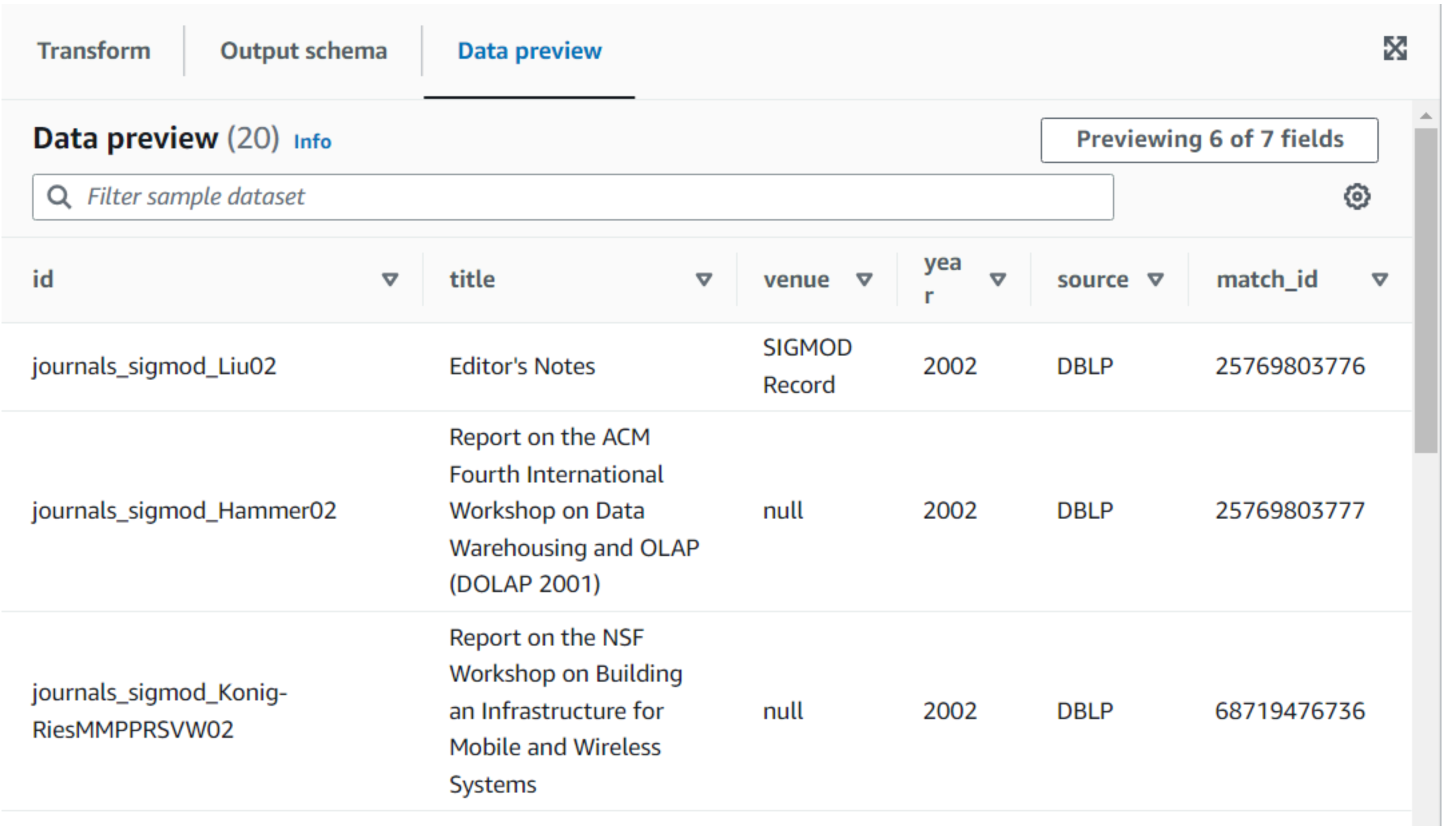
☐

Using the Record Matching transform to invoke an existing data classification transform

This transform invokes an existing Record Matching machine learning data classification transform.

The transform evaluates the current data against the trained model based on labels. A column "match_id" is added to assign each row to a group of items that are considered equivalent based on the algorithm training. For more information, see [Record matching with Lake Formation FindMatches](#).

> **Note**
>
> The version of AWS Glue used by the visual job must match the version that AWS Glue used to create the Record Matching transform.

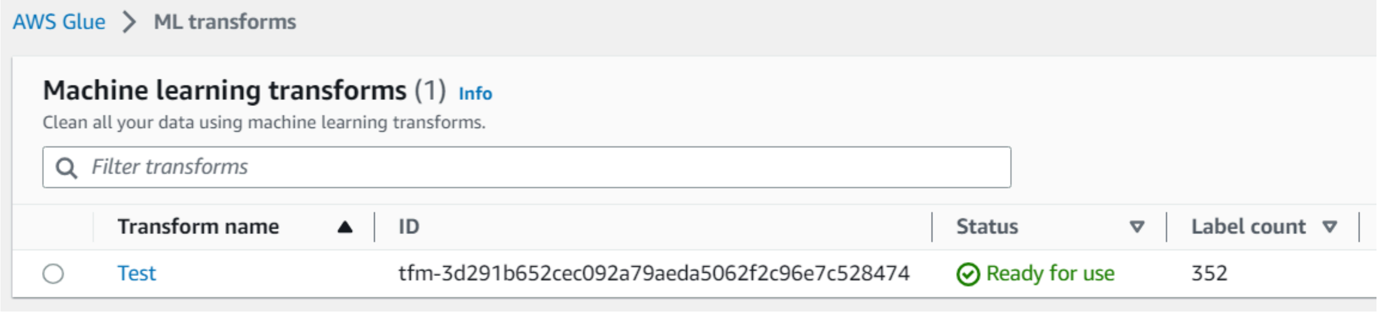


id	title	venue	year	source	match_id
journals_sigmod_Liu02	Editor's Notes	SIGMOD Record	2002	DBLP	25769803776
journals_sigmod_Hammer02	Report on the ACM Fourth International Workshop on Data Warehousing and OLAP (DOLAP 2001)	null	2002	DBLP	25769803777
journals_sigmod_Konig-RiesMMPPRSVW02	Report on the NSF Workshop on Building an Infrastructure for Mobile and Wireless Systems	null	2002	DBLP	68719476736

To add a Record Matching transform node to your job diagram

1. Open the Resource panel, and then choose **Record Matching** to add a new transform to your job diagram. The node selected at the time of adding the node will be its parent.

2. In the node properties panel, you can enter a name for the node in the job diagram. If a node parent isn't already selected, choose a node from the **Node parents** list to use as the input source for the transform.
3. On the **Transform** tab, enter the ID taken from the **Machine learning transforms** page:



AWS Glue > ML transforms

Machine learning transforms (1) Info

Clean all your data using machine learning transforms.

Filter transforms

	Transform name ▲	ID	Status ▽	Label count ▽
○	Test	tfm-3d291b652cec092a79aeda5062f2c96e7c528474	✓ Ready for use	352

4. (Optional) On the **Transform** tab, you can check the option to add the confidence scores. At the cost of extra computing, the model will estimate a confidence score for each match as an additional column.

Removing null rows

This transform removes from the dataset rows that have all columns as null. In addition, you can extend this criteria to include empty fields, so as to keep rows where at least one column is non empty.

To add a Remove Null Rows transform node to your job diagram

1. Open the Resource panel, and then choose **Remove Null Rows** to add a new transform to your job diagram. The node selected at the time of adding the node will be its parent.
2. In the node properties panel, you can enter a name for the node in the job diagram. If a node parent isn't already selected, choose a node from the **Node parents** list to use as the input source for the transform.
3. (Optional) On the **Transform** tab, check the **Extended** option if you want to require rows not just to not be null but also not empty, this way empty strings, arrays or maps will be considered nulls for the purpose of this transform.

Parsing a string column containing JSON data

This transform parses a string column containing JSON data and convert it to a struct or an array column, depending if the JSON is an object or an array, respectively. Optionally you can keep both the parsed and original column.

The JSON schema can be provided or inferred (in the case of JSON objects), with optional sampling.

To add a Parse JSON Column transform node to your job diagram

1. Open the Resource panel, and then choose **Parse JSON Column** to add a new transform to your job diagram. The node selected at the time of adding the node will be its parent.
2. In the node properties panel, you can enter a name for the node in the job diagram. If a node parent isn't already selected, choose a node from the **Node parents** list to use as the input source for the transform.
3. On the **Transform** tab, select the column containing the JSON string.
4. (Optional) On the **Transform** tab, enter the schema that the JSON data follows using SQL syntax, for instance: "field1 STRING, field2 INT" in the case of an object or "ARRAY<STRING>" in the case of an array.

   If the case of an array the schema is required but in the case of an object, if the schema is not specified it will be inferred using the data. To reduce the impact of inferring the schema (especially on a large dataset), you can avoid reading the whole data twice by entering a **Ratio of samples to use to infer schema**. If the value is lower than 1, the corresponding ratio of random samples is used to infer the schema. If the data is reliable and the object is consistent between rows, you can use a small ratio such as 0.1 to improve performance.
5. (Optional) On the **Transform** tab, you can enter a new column name if you want to keep both the original string column and the parsed column.

Extracting a JSON path

This transform extracts new columns from a JSON string column. This transform is useful when you only need a few data elements and don't want to import the entire JSON content into the table schema.

To add an Extract JSON Path transform node to your job diagram

1. Open the Resource panel, and then choose **Extract JSON Path** to add a new transform to your job diagram. The node selected at the time of adding the node will be its parent.
2. In the node properties panel, you can enter a name for the node in the job diagram. If a node parent isn't already selected, choose a node from the **Node parents** list to use as the input source for the transform.
3. On the **Transform** tab, select the column containing the JSON string. Enter one of more JSON path expressions separated by commas, each one referencing how to extract a value out of the JSON array or object. For instance, if the JSON column contained an objects with properties "prop_1" and "prop2" you could extract both specifying their names "prop_1, prop_2".

   If the JSON field has special characters, for instance to extract the property from this JSON `{"a. a": 1}` you could use the path `$['a. a']`. The exception is the comma because it is reserved to separate paths. Then enter the corresponding column names for each path, separated by commas.
4. (Optional) On the **Transform** tab, you can check to drop the JSON column once extracted, this makes sense when you don't need the rest of the JSON data once you have extracted the parts you need.

Extracting string fragments using a regular expression

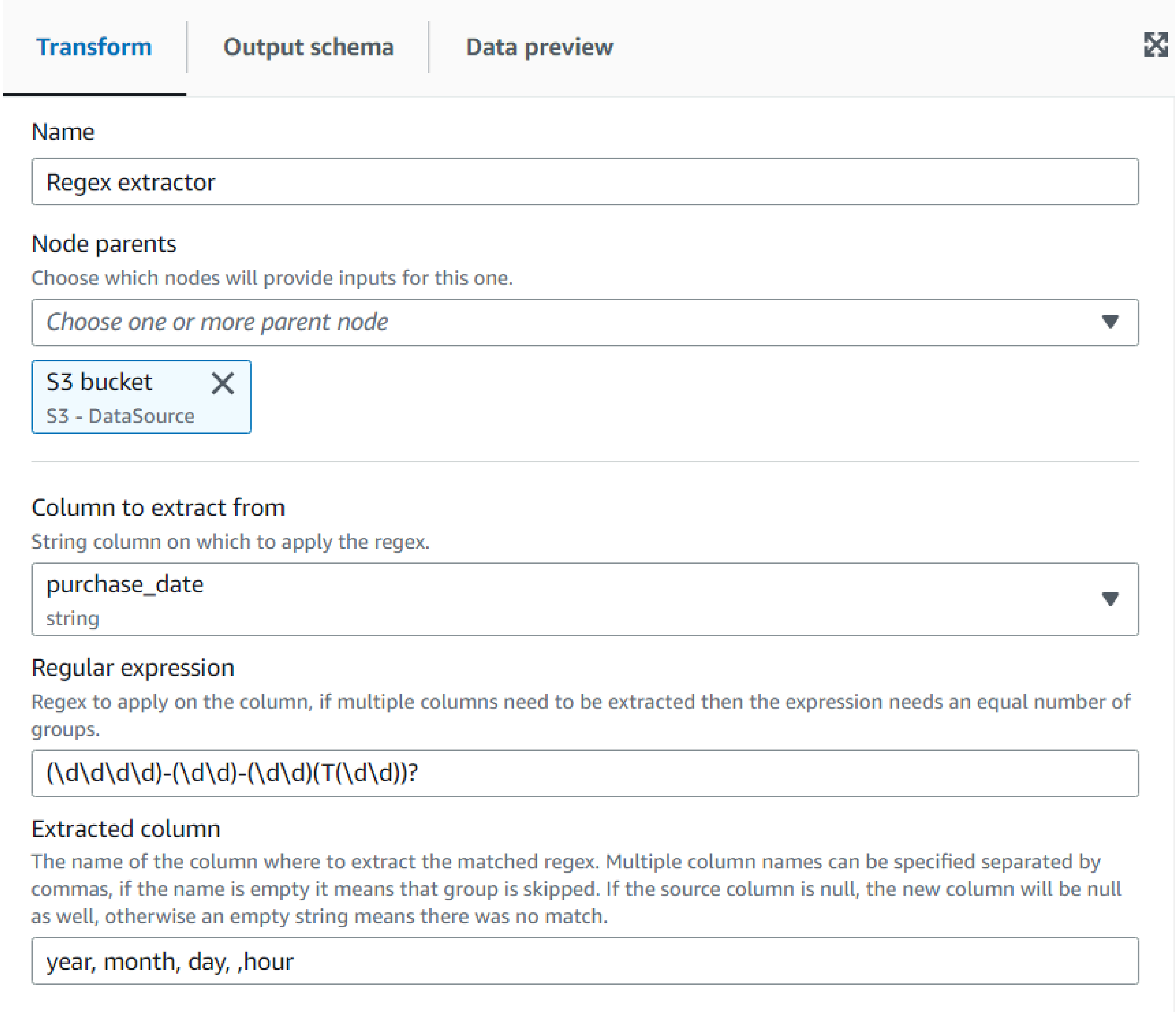
This transform extracts string fragments using a regular expression and creates a new column out of it, or multiple columns if using regex groups.

To add a Regex Extractor transform node to your job diagram

1. Open the Resource panel, and then choose **Regex Extractor** to add a new transform to your job diagram. The node selected at the time of adding the node will be its parent.
2. In the node properties panel, you can enter a name for the node in the job diagram. If a node parent isn't already selected, choose a node from the **Node parents** list to use as the input source for the transform.
3. On the **Transform** tab, enter the regular expression and the column on which it needs to be applied. Then enter the name of the new column on which to store the matching string. The new column will be null only if the source column is null, if the regex doesn’t match the column will be empty.

If the regex uses groups, there has be a corresponding column name separated by comma but you can skip groups by leaving the column name empty.

For example, if you have a column "purchase_date" with a string using both long and short ISO date formats, then you want to extract the year, month, day and hour, when available. Notice the hour group is optional, otherwise in the rows where not available, all the extracted groups would be empty strings (because the regex didn't match). In this case, we don't want the group to make the time optional but the inner one, so we leave the name empty and it doesn't get extracted (that group would include the T character).



Name

Regex extractor

Node parents

Choose which nodes will provide inputs for this one.

Choose one or more parent node

S3 bucket

S3 - DataSource

Column to extract from

String column on which to apply the regex.

purchase_date

string

Regular expression

Regex to apply on the column, if multiple columns need to be extracted then the expression needs an equal number of groups.

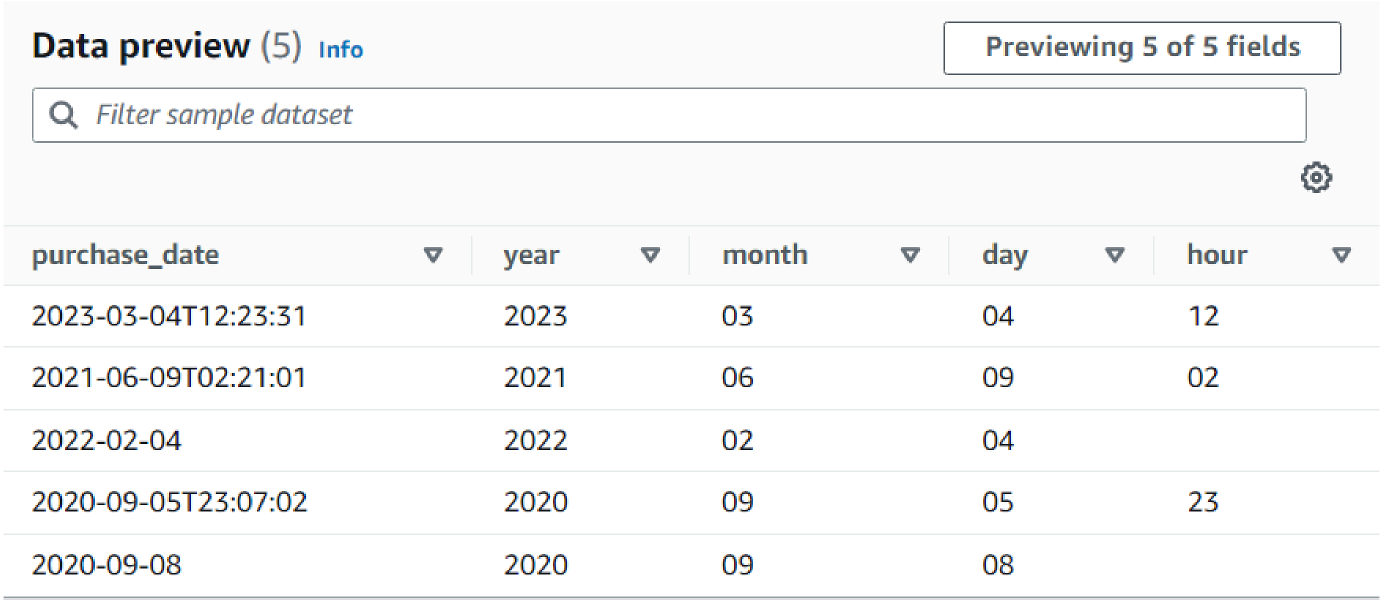
(\d\d\d\d)-(\d\d)-(\d\d)(T(\d\d))?

Extracted column

The name of the column where to extract the matched regex. Multiple column names can be specified separated by commas, if the name is empty it means that group is skipped. If the source column is null, the new column will be null as well, otherwise an empty string means there was no match.

year, month, day, ,hour

Resulting in the data preview:



purchase_date	year	month	day	hour
2023-03-04T12:23:31	2023	03	04	12
2021-06-09T02:21:01	2021	06	09	02
2022-02-04	2022	02	04	
2020-09-05T23:07:02	2020	09	05	23
2020-09-08	2020	09	08	

Creating a custom transformation

If you need to perform more complicated transformations on your data, or want to add data property keys to the dataset, you can add a **Custom code** transform to your job diagram. The Custom code node allows you to enter a script that performs the transformation.

When using custom code, you must use a schema editor to indicate the changes made to the output through the custom code. When editing the schema, you can perform the following actions:

- Add or remove data property keys
- Change the data type of data property keys
- Change the name of data property keys
- Restructure a nested property key

You must use a *SelectFromCollection* transform to choose a single `DynamicFrame` from the result of your Custom transform node before you can send the output to a target location.

Use the following tasks to add a custom transform node to your job diagram.

Adding a custom code transform node to the job diagram

To add a custom transform node to your job diagram

1. (Optional) Open the Resource panel and then choose **Custom transform** to add a custom transform to your job diagram.
2. On the **Node properties** tab, enter a name for the node in the job diagram. If a node parent is not already selected, or if you want multiple inputs for the custom transform, then choose a node from the **Node parents** list to use as the input source for the transform.

Entering code for the custom transform node

You can type or copy code into an input field. The job uses this code to perform the data transformation. You can provide a code snippet in either Python or Scala. The code should take one or more `DynamicFrames` as input and returns a collection of `DynamicFrames`.

To enter the script for a custom transform node

1. With the custom transform node selected in the job diagram, choose the **Transform** tab.
2. In the text entry field under the heading **Code block**, paste or enter the code for the transformation. The code that you use must match the language specified for the job on the **Job details** tab.

   When referring to the input nodes in your code, AWS Glue Studio names the `DynamicFrames` returned by the job diagram nodes sequentially based on the order of creation. Use one of the following naming methods in your code:

   - Classic code generation – Use functional names to refer to the nodes in your job diagram.
     - Data source nodes: `DataSource0`, `DataSource1`, `DataSource2`, and so on.
     - Transform nodes: `Transform0`, `Transform1`, `Transform2`, and so on.
   - New code generation – Use the name specified on the **Node properties** tab of a node, appended with '_node1', '_node2', and so on. For example, `S3bucket_node1`, `ApplyMapping_node2`, `S3bucket_node2`, `MyCustomNodeName_node1`.

   For more information about the new code generator, see Script code generation.

The following examples show the format of the code to enter in the code box:

Python

The following example takes the first `DynamicFrame` received, converts it to a `DataFrame` to apply the native filter method (keeping only records that have over 1000 votes), then converts it back to a `DynamicFrame` before returning it.

```
def FilterHighVoteCounts (glueContext, dfc) -> DynamicFrameCollection:
    df = dfc.select(list(dfc.keys())[0]).toDF()
    df_filtered = df.filter(df["vote_count"] > 1000)
    dyf_filtered = DynamicFrame.fromDF(df_filtered, glueContext, "filter_votes")
    return(DynamicFrameCollection({"CustomTransform0": dyf_filtered}, glueContext))
```

Scala

The following example takes the first `DynamicFrame` received, converts it to a `DataFrame` to apply the native filter method (keeping only records that have over 1000 votes), then converts it back to a `DynamicFrame` before returning it.

```
object FilterHighVoteCounts {
  def execute(glueContext : GlueContext, input : Seq[DynamicFrame]) :
 Seq[DynamicFrame] = {
    val frame = input(0).toDF()
    val filtered = DynamicFrame(frame.filter(frame("vote_count") > 1000),
 glueContext)
    Seq(filtered)
  }
}
```

Editing the schema in a custom transform node

When you use a custom transform node, AWS Glue Studio cannot automatically infer the output schemas created by the transform. You use the schema editor to describe the schema changes implemented by the custom transform code.

A custom code node can have any number of parent nodes, each providing a `DynamicFrame` as input for your custom code. A custom code node returns a collection of `DynamicFrames`. Each `DynamicFrame` that is used as input has an associated schema. You must add a schema that describes each `DynamicFrame` returned by the custom code node.

> **Note**
>
> When you set your own schema on a custom transform, AWS Glue Studio does not inherit schemas from previous nodes.To update the schema, select the Custom transform node, then choose the Data preview tab. Once the preview is generated, choose 'Use Preview Schema'. The schema will then be replaced by the schema using the preview data.

To edit the output schemas for a custom transform node

1. With the custom transform node selected in the job diagram, in the node details panel, choose the **Output schema** tab.
2. Choose **Edit** to make changes to the schema.

   If you have nested data property keys, such as an array or object, you can choose the **Expand-Rows** icon (↧) on the top right of each schema panel to expand the list of child data property keys. After you choose this icon, it changes to the **Collapse-Rows** icon (⇞), which you can choose to collapse the list of child property keys.
3. Modify the schema using the following actions in the section on the right side of the page:

   - To rename a property key, place the cursor in the **Key** text box for the property key, then enter the new name.
   - To change the data type for a property key, use the list to choose the new data type for the property key.
   - To add a new top-level property key to the schema, choose the **Overflow** (⋯) icon to the left of the **Cancel** button, and then choose **Add root key**.
   - To add a child property key to the schema, choose the **Add-Key** icon ↳ associa with the parent key. Enter a name for the child key and choose the data type.
   - To remove a property key from the schema, choose the **Remove** icon (🗑) to the far right of the key name.

4. If your custom transform code uses multiple `DynamicFrames`, you can add additional output schemas.

    - To add a new, empty schema, choose the **Overflow** (⋯) icon, and then choose **Add output schema**.
    - To copy an existing schema to a new output schema, make sure the schema you want to copy is displayed in the schema selector. Choose the **Overflow** (⋯) icon, and then choose **Duplicate**.

    If you want to remove an output schema, make sure the schema you want to copy is displayed in the schema selector. Choose the **Overflow** (⋯) icon, and then choose **Delete**.
5. Add new root keys to the new schema or edit the duplicated keys.
6. When you are modifying the output schemas, choose the **Apply** button to save your changes and exit the schema editor.

    If you do not want to save your changes, choose the **Cancel** button.

Configure the custom transform output

A custom code transform returns a collection of `DynamicFrames`, even if there is only one `DynamicFrame` in the result set.

To process the output from a custom transform node

1. Add a *SelectFromCollection* transform node, which has the custom transform node as its parent node. Update this transform to indicate which dataset you want to use. See Using SelectFromCollection to choose which dataset to keep for more information.
2. Add additional *SelectFromCollection* transforms to the job diagram if you want to use additional `DynamicFrames` produced by the custom transform node.

    Consider a scenario in which you add a custom transform node to split a flight dataset into multiple datasets, but duplicate some of the identifying property keys in each output schema, such as the flight date or flight number. You add a *SelectFromCollection* transform node for each output schema, with the custom transform node as its parent.

3. (Optional) You can then use each *SelectFromCollection* transform node as input for other nodes in the job, or as a parent for a data target node.

Transform data with custom visual transforms

Custom visual transforms allow you to create transforms and make them available for use in AWS Glue Studio jobs. Custom visual transforms enable ETL developers, who may not be familiar with coding, to search and use a growing library of transforms using the AWS Glue Studio interface.

You can create a custom visual transform, then upload it to Amazon S3 to make available for use through the visual editor in AWS Glue Studio to work with these jobs.

Topics

- Getting started with custom visual transforms
- Step 1. Create a JSON config file
- Step 2. Implement the transform logic
- Step 3. Validate and troubleshoot custom visual transforms in AWS Glue Studio
- Step 4. Update custom visual transforms as needed
- Step 5. Use custom visual transforms in AWS Glue Studio
- Usage examples
- Examples of custom visual scripts
- Video

Getting started with custom visual transforms

To create a custom visual transform, you go through the following steps.

- Step 1. Create a JSON config file
- Step 2. Implement the transform logic
- Step 3. Validate the custom visual transform
- Step 4. Update the custom visual transform as needed
- Step 5. Use the custom visual transform in AWS Glue Studio

Get started by setting up the Amazon S3 bucket and continue to **Step 1. Create a JSON config file.**

Prerequisites

Customer-supplied transforms reside within a customer AWS account. That account owns the transforms and therefore has all permissions to view (search and use), edit, or delete them.

In order to use a custom transform in AWS Glue Studio, you will need to create and upload two files to the Amazon S3 assets bucket in that AWS account:

- **Python file** – contains the transform function
- **JSON file –** describes the transform. This is also known as the config file that is required to define the transform.

In order to pair the files together, use the same name for both. For example:

- myTransform.json
- myTransform.py

Optionally, you can give your custom visual transform a custom icon by providing a **SVG file** containing the icon. In order to pair the files together, use the same name for the icon:

- myTransform.svg

AWS Glue Studio will automatically match them using their respective file names. File names cannot be the same for any existing module.

Recommended convention for transform file name

AWS Glue Studio will import your file as module (for example, `import myTransform`) in your job script. Therefore, your file name must follow the same naming rules set for python variable names (identifiers). Specifically, they must start with either a letter or an underscore and then be composed entirely of letters, digits, and/or underscores.

> **Note**
>
> Ensure your transform file name is not conflicting with existing loaded python modules (for example, `sys, array, copy` etc.) to avoid unexpected runtime issues.

Setting up the Amazon S3 bucket

Transforms you create are stored in Amazon S3 and is owned by your AWS account. You create new custom visual transforms by simply uploading files (json and py) to the Amazon S3 assets folder where all job scripts are currently stored (for example, `s3://aws-glue-assets-<accountid>-<region>/transforms`). If using a custom icon, upload it as well. By default, AWS Glue Studio will read all .json files from the /transforms folder in the same S3 bucket.

Step 1. Create a JSON config file

A JSON config file is required to define and describe your custom visual transform. The schema for the config file is as follows.

JSON file structure

Fields

- `name: string` – (required) the transform system name used to identify transforms. Follow the same naming rules set for python variable names (identifiers). Specifically, they must start with either a letter or an underscore and then be composed entirely of letters, digits, and/or underscores.
- `displayName: string` – (optional) the name of the transform displayed in the AWS Glue Studio visual job editor. If no `displayName` is specified, the `name` is used as the name of the transform in AWS Glue Studio.
- `description: string` – (optional) the transform description is displayed in AWS Glue Studio and is searchable.
- `functionName: string` – (required) the Python function name is used to identify the function to call in the Python script.
- `path: string` – (optional) the full Amazon S3 path to the Python source file. If not specified, AWS Glue uses file name matching to pair the .json and .py files together. For example, the name of the JSON file, `myTransform.json`, will be paired to the Python file, `myTransform.py`, on the same Amazon S3 location.
- `parameters: Array of TransformParameter object` – (optional) the list of parameters to be displayed when you configure them in the AWS Glue Studio visual editor.

TransformParameter fields

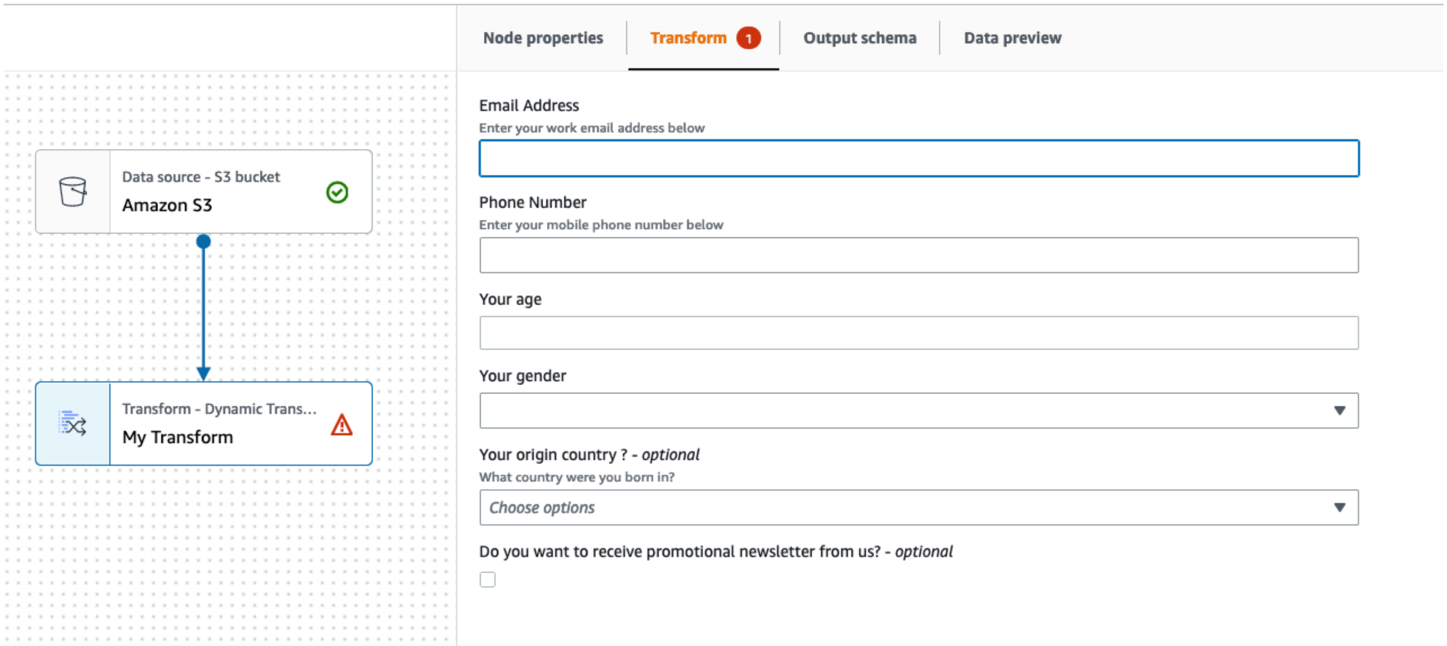
- `name: string` – (required) the parameter name that will be passed to the python function as a named argument in the job script. Follow the same naming rules set for python variable names (identifiers). Specifically, they must start with either a letter or an underscore and then be composed entirely of letters, digits, and/or underscores.
- `displayName: string` – (optional) the name of the transform displayed in the AWS Glue Studio visual job editor. If no `displayName` is specified, the `name` is used as the name of the transform in AWS Glue Studio.
- `type: string` – (required) the parameter type accepting common Python data types. Valid values: 'str' | 'int' | 'float' | 'list' | 'bool'.
- `isOptional: boolean` – (optional) determines whether the parameter is optional. By default all parameters are required.
- `description: string` — (optional) description is displayed in AWS Glue Studio to help the user configure the transform parameter.
- `validationType: string` – (optional) defines the way this parameter is validated. Currently, it only supports regular expressions. By default, the validation type is set to `RegularExpression`.
- `validationRule: string` – (optional) regular expression used to validate form input before submit when `validationType` is set to `RegularExpression`. Regular expression syntax must be compatible with RegExp Ecmascript specifications.
- `validationMessage: string` – (optional) the message to display when validation fails.
- `listOptions: An array of TransformParameterListOption object` OR a `string` or the string value ‘column’ – (optional) options to display in Select or Multiselect UI control. Accepting a list of comma separated value or a strongly type JSON object of type `TransformParameterListOption`. It can also dynamically populate the list of columns from the parent node schema by specifying the string value “column”.
- `listType: string` – (optional) Define options types for type = 'list'. Valid values: 'str' | 'int' | 'float' | 'list' | 'bool'. Parameter type accepting common python data types.

TransformParameterListOption fields

- `value: string | int | float | bool` – (required) option value.
- `label: string` – (optional) option label displayed in the select dropdown.

Transform parameters in AWS Glue Studio

By default, parameters are required unless mark as `isOptional` in the .json file. In AWS Glue Studio, parameters are displayed in the **Transform** tab. The example shows user-defined parameters such as Email Address, Phone Number, Your age, Your gender and Your origin country.



Email Address
Enter your work email address below

Phone Number
Enter your mobile phone number below

Your age

Your gender

Your origin country ? - *optional*
What country were you born in?
Choose options

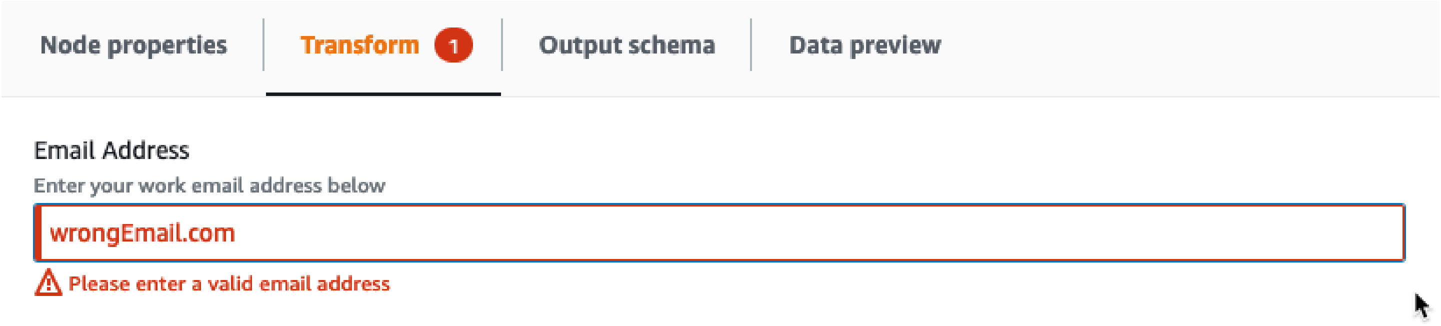
Do you want to receive promotional newsletter from us? - *optional*
☐

You can enforce some validations in AWS Glue Studio using regular expressions in the json file by specifying the `validationRule` parameter and specifying a validation message in `validationMessage`.

```
      "validationRule": "^\\(?(\\d{3})\\)?[- ]?(\\d{3})[- ]?(\\d{4})$",
      "validationMessage": "Please enter a valid US number"
```

> **Note**
>
> Since validation occurs in the browser, your regular expression syntax must be compatible with RegExp Ecmascript specifications. Python syntax is not supported for these regular expressions.

Adding validation will prevent the user from saving the job with incorrect user input. AWS Glue Studio displays the validation message as displayed in the example:



Email Address

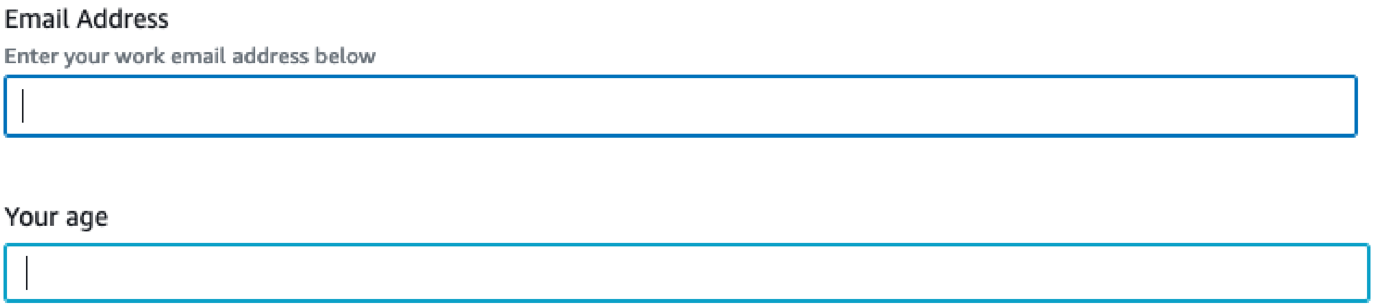
Enter your work email address below

wrongEmail.com

Please enter a valid email address

Parameters are displayed in AWS Glue Studio based on the parameter configuration.

- When `type` is any of the following: `str`, `int` or `float`, a text input field is displayed. For example, the screenshot shows input fields for 'Email Address' and 'Your age' parameters.

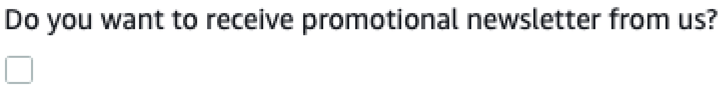


Email Address

Enter your work email address below

Your age

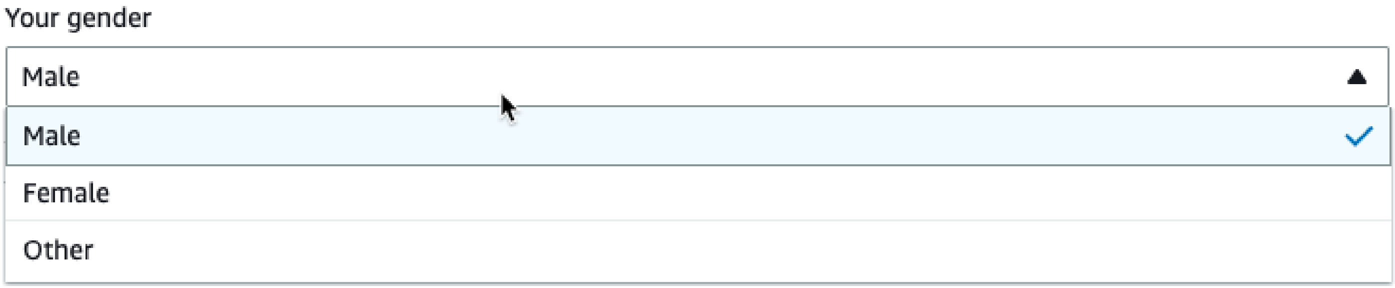
- When `type` is `bool`, a checkbox is displayed.



Do you want to receive promotional newsletter from us?

☐

- When `type` is `str` and `listOptions` is provided, a single select list is displayed.



Your gender

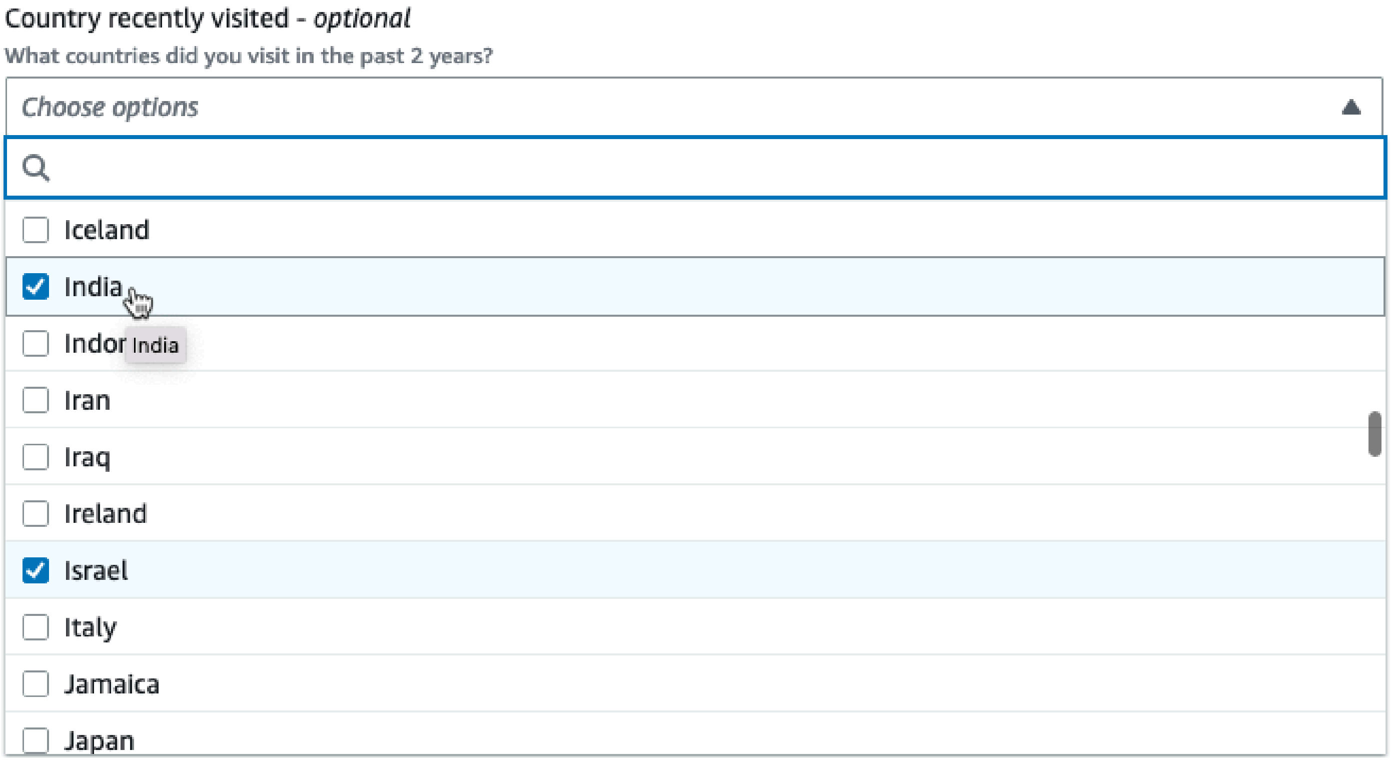
Male

Male ✓

Female

Other

- When `type` is `list` and `listOptions` and `listType` are provided, a multi-select list is displayed.



Country recently visited - *optional*

What countries did you visit in the past 2 years?

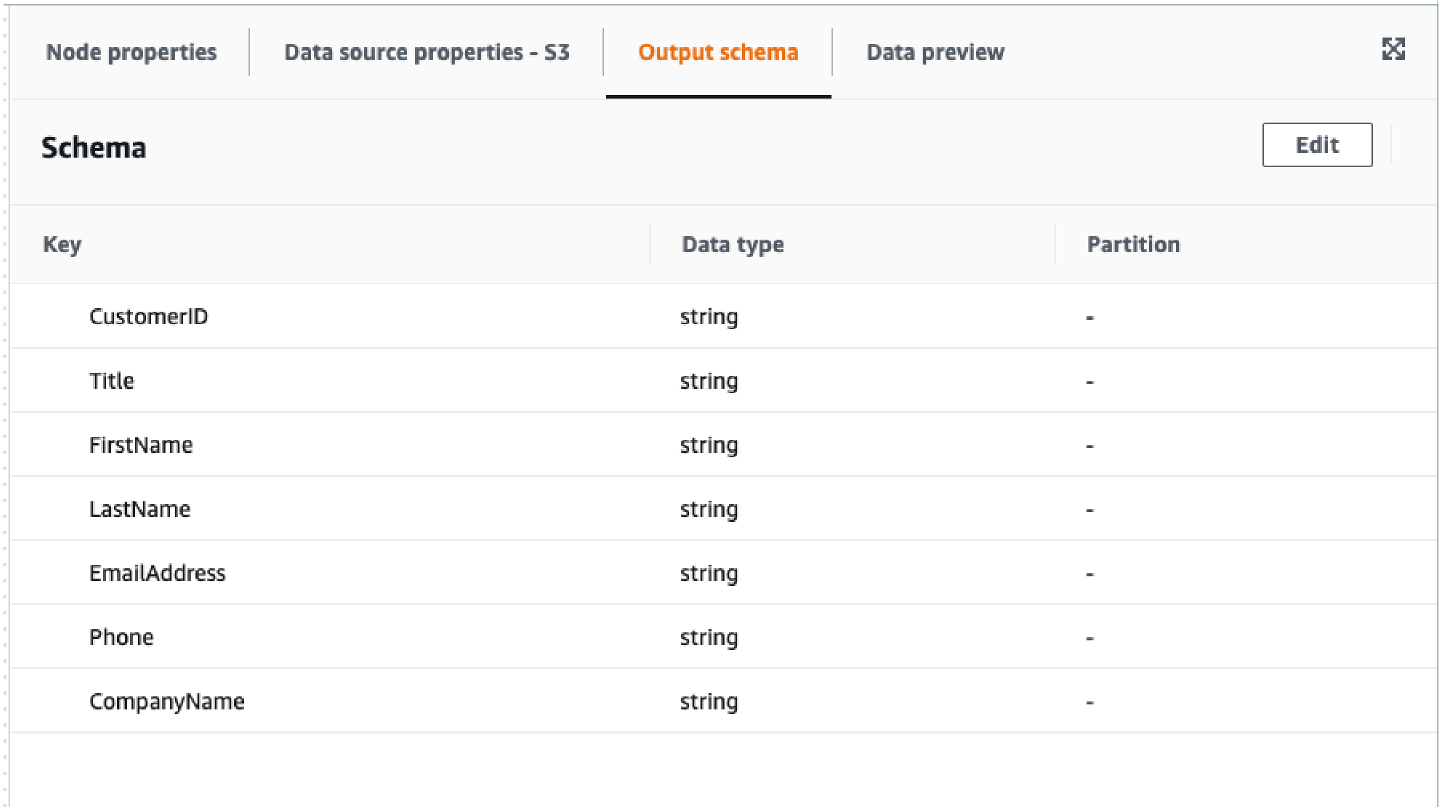
Choose options

- ☐ Iceland
- ☑ India
- ☐ Indor India
- ☐ Iran
- ☐ Iraq
- ☐ Ireland
- ☑ Israel
- ☐ Italy
- ☐ Jamaica
- ☐ Japan

Displaying a column selector as parameter

If the configuration requires the user to choose a column from the schema, you can display a column selector so the user isn't required to type the column name. By setting the `listOptions` field to '"column", AWS Glue Studio dynamically displays a column selector based on the parent node output schema. AWS Glue Studio can display either a single or multiple column selector.

The following example uses the schema:

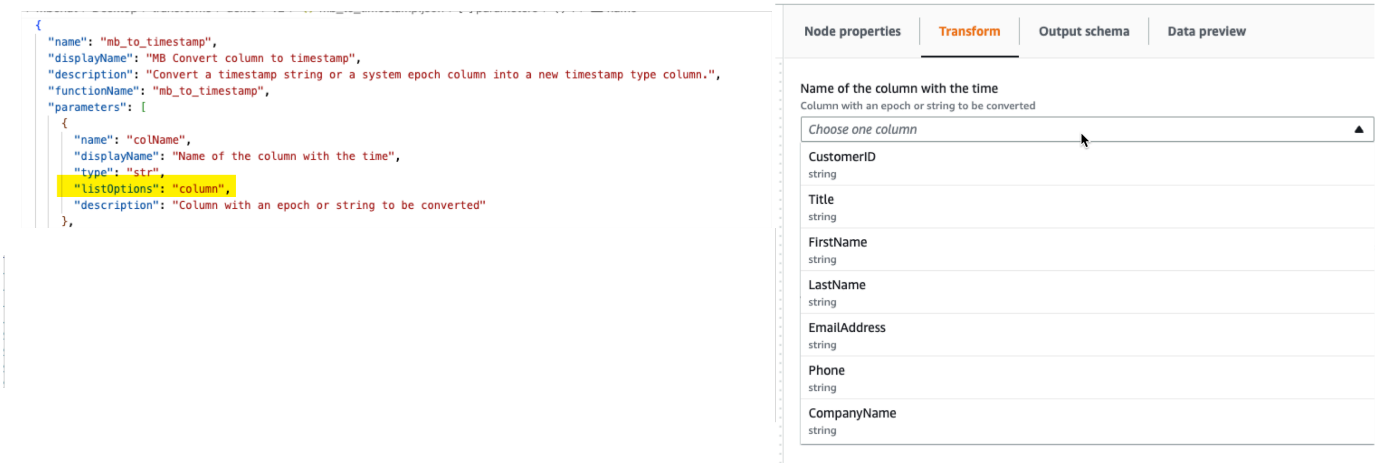

Node properties
Data source properties - S3
Output schema
Data preview
Edit

Schema

Key	Data type	Partition
CustomerID	string	-
Title	string	-
FirstName	string	-
LastName	string	-
EmailAddress	string	-
Phone	string	-
CompanyName	string	-

To define your Custom Visual Transform parameter to display a single column:

1. In your JSON file, for the `parameters` object, set the `listOptions` value to "column". This allows a user to choose a column from a pick list in AWS Glue Studio.



```
{
  "name": "mb_to_timestamp",
  "displayName": "MB Convert column to timestamp",
  "description": "Convert a timestamp string or a system epoch column into a new timestamp type column.",
  "functionName": "mb_to_timestamp",
  "parameters": [
    {
      "name": "colName",
      "displayName": "Name of the column with the time",
      "type": "str",
      "listOptions": "column",
      "description": "Column with an epoch or string to be converted"
    },
```

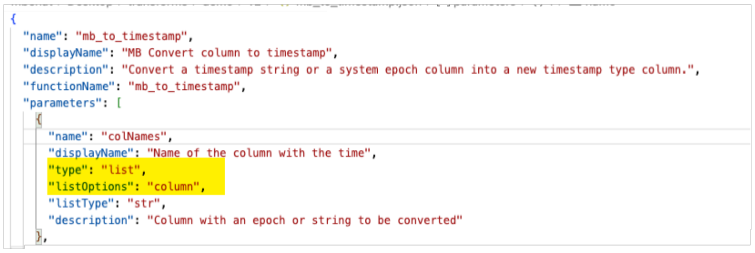
Node properties
Transform
Output schema
Data preview

Name of the column with the time
Column with an epoch or string to be converted
Choose one column

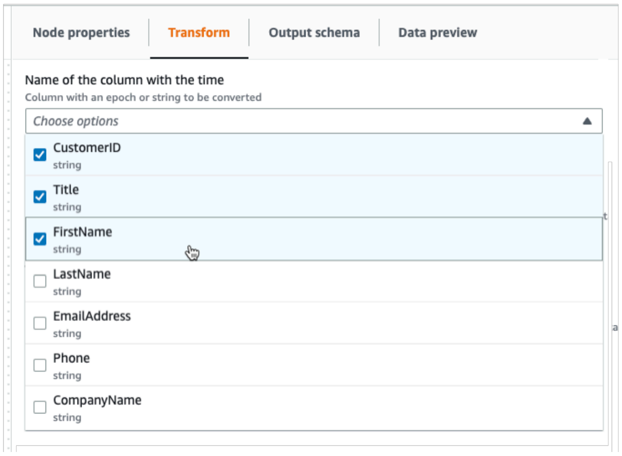
CustomerID
string
Title
string
FirstName
string
LastName
string
EmailAddress
string
Phone
string
CompanyName
string

2. You can also allow multiple columns selection by defining the parameter as:

   - `listOptions: "column"`
   - `type: "list"`



```
{
  "name": "mb_to_timestamp",
  "displayName": "MB Convert column to timestamp",
  "description": "Convert a timestamp string or a system epoch column into a new timestamp type column.",
  "functionName": "mb_to_timestamp",
  "parameters": [
    {
      "name": "colNames",
      "displayName": "Name of the column with the time",
      "type": "list",
      "listOptions": "column",
      "listType": "str",
      "description": "Column with an epoch or string to be converted"
    },
```



Name of the column with the time

Column with an epoch or string to be converted

Choose options

- ☑ CustomerID string
- ☑ Title string
- ☑ FirstName string
- ☐ LastName string
- ☐ EmailAddress string
- ☐ Phone string
- ☐ CompanyName string

Step 2. Implement the transform logic

> **Note**
>
> Custom visual transforms only support Python scripts. Scala is not supported.

To add the code that implements the function defined by the .json config file, it is recommended to place the Python file in the same location as the .json file, with the same name but with the ".py" extension. AWS Glue Studio automatically pairs the .json and .py files so that you don't need to specify the path of the Python file in the config file.

In the Python file, add the declared function, with the named parameters configured and register it to be used in `DynamicFrame`. The following is an example of a Python file:

```
from awsglue import DynamicFrame

# self refers to the DynamicFrame to transform,
# the parameter names must match the ones defined in the config
# if it's optional, need to provide a default value
def myTransform(self, email, phone, age=None, gender="",
                      country="", promotion=False):
   resulting_dynf = # do some transformation on self
   return resulting_dynf

DynamicFrame.myTransform = myTransform
```

It is recommended to use an AWS Glue notebook for the quickest way to develop and test the python code. See [Getting started with notebooks in AWS Glue Studio](#).

To illustrate how to implement the transform logic, the custom visual transform in the example below is a transform to filter incoming data to keep only the data related to a specific US state. The .json file contains the parameter for `functionName` as `custom_filter_state` and two arguments ("state" and "colName" with type "str").

The example config .json file is:

```
{
"name": "custom_filter_state",
"displayName": "Filter State",
"description": "A simple example to filter the data to keep only the state indicated.",
"functionName": "custom_filter_state",
"parameters": [
   {
    "name": "colName",
    "displayName": "Column name",
    "type": "str",
    "description": "Name of the column in the data that holds the state postal code"
   },
   {
    "name": "state",
    "displayName": "State postal code",
    "type": "str",
    "description": "The postal code of the state whole rows to keep"
   }
  ]
}
```

To implement the companion script in Python

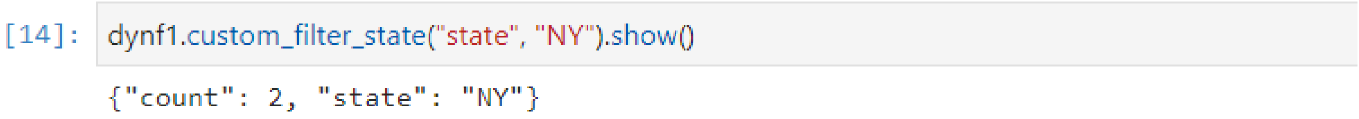
1. Start a AWS Glue notebook and run the initial cell provided for the session to be started. Running the initial cell creates the basic components required.
2. Create a function that performs the filtering as describe in the example and register it on `DynamicFrame`. Copy the code below and paste into a cell in the AWS Glue notebook.

```
from awsglue import DynamicFrame

def custom_filter_state(self, colName, state):
    return self.filter(lambda row: row[colName] == state)

DynamicFrame.custom_filter_state = custom_filter_state
```

3. Create or load sample data to test the code in the same cell or a new cell. If you add the sample data in a new cell, don't forget to run the cell. For example:

```
# A few of rows of sample data to test
data_sample = [
    {"state": "CA", "count": 4},
    {"state": "NY", "count": 2},
    {"state": "WA", "count": 3}
]
df1 = glueContext.sparkSession.sparkContext.parallelize(data_sample).toDF()
dynf1 = DynamicFrame.fromDF(df1, glueContext, None)
```

4. Test to validate the “custom_filter_state” with different arguments:



```
[14]: dynf1.custom_filter_state("state", "NY").show()
      {"count": 2, "state": "NY"}
```

5. After running several tests, save the code with the .py extension and name the .py file with a name that mirrors the .json file name. The .py and .json files should be in the same transform folder.

   Copy the following code and paste it to a file and rename it with a .py file extension.

```
from awsglue import DynamicFrame

def custom_filter_state(self, colName, state):
    return self.filter(lambda row: row[colName] == state)

DynamicFrame.custom_filter_state = custom_filter_state
```

6. In AWS Glue Studio, open a visual job and add the transform to the job by selecting it from the list of available **Transforms**.

   To reuse this transform in a Python script code, add the Amazon S3 path to the .py file in the job under “Referenced files path” and in the script, import the name of the python file (without the extension) by adding it to the top of the file. For example: `import` <name of the file (without the extension)>

Step 3. Validate and troubleshoot custom visual transforms in AWS Glue Studio

AWS Glue Studio validates the JSON config file before custom visual transforms are loaded into AWS Glue Studio. Validation includes:

- Presence of required fields
- JSON format validation
- Incorrect or invalid parameters
- Presence of both the .py and .json files in the same Amazon S3 path
- Matching filenames for the .py and .json

If validation succeeds, the transform is listed in the list of available **Actions** in the visual editor. If a custom icon has been provided, it should be visible beside the **Action**.

If validation fails, AWS Glue Studio does not load the custom visual transform.

Step 4. Update custom visual transforms as needed

Once created and used, the transform script can be updated as long as the transform follows the corresponding json definition:

- The name used when assigning to DynamicFrame much match the json `functionName`.
- The function arguments must be defined in the json file as described in Step 1. Create a JSON config file .
- The Amazon S3 path of the Python file cannot change, since the jobs depend directly on it.

> **Note**
>
> If any updates need to be made, ensure the script and the .json file are consistently updated and any visual jobs are correctly saved again with the new transform. If visual jobs are not saved after the updates were made, the updates will not be applied and validated. If the Python script file is renamed or not placed next to the .json file, then you need to specify the full path in the .json file.

Custom icon

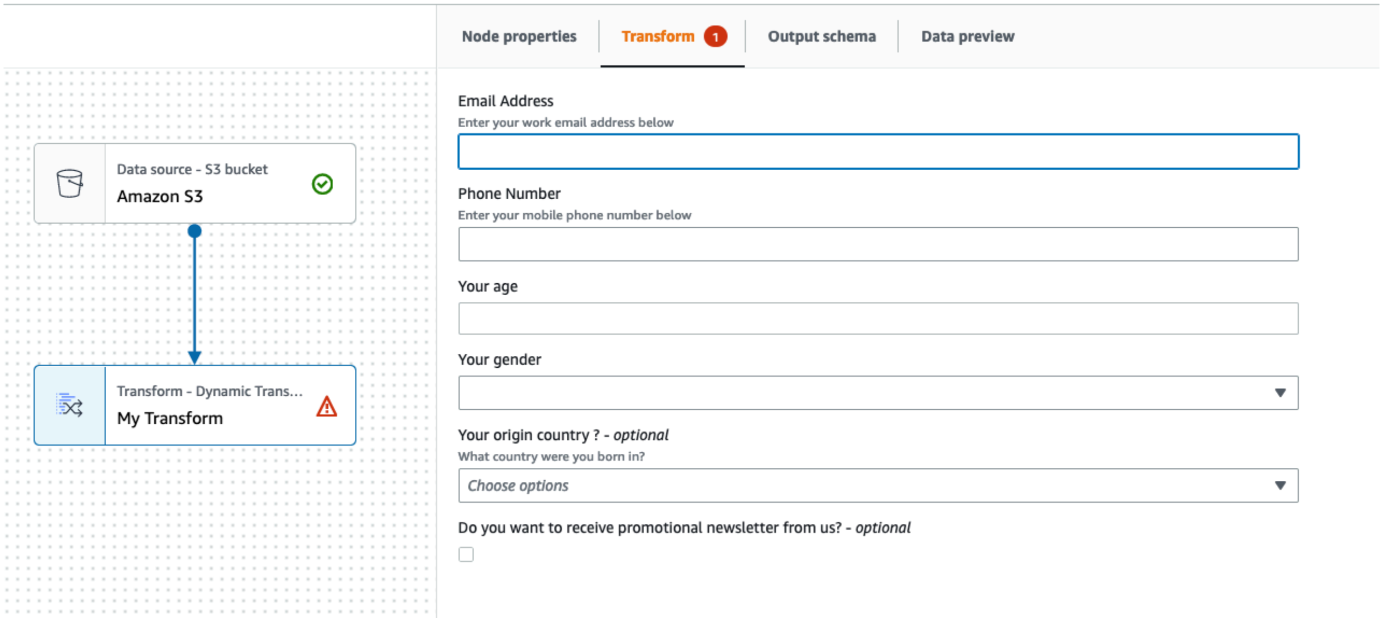
If you determine the default icon for your **Action** does not visually distinguish it as part of your workflows, you can provide a custom icon, as described in the section called " Getting started with custom visual transforms ". You can update the icon by updating the corresponding SVG hosted in Amazon S3.

For best results, design your image to be viewed at 32x32px following guidelines from the Cloudscape Design System. For more information about Cloudscape guidelines, see The Cloudscape documentation

Step 5. Use custom visual transforms in AWS Glue Studio

To use a custom visual transform in AWS Glue Studio, you upload the config and source files, then select the transform from the **Action** menu. Any parameters that need values or input are available to you in the **Transform** tab.

1. Upload the two files (Python source file and JSON config file) to the Amazon S3 assets folder where the job scripts are stored. By default, AWS Glue pulls all JSON files from the **/transforms** folder within the same Amazon S3 bucket.
2. From the **Action** menu, choose the custom visual transform. It is named with the transform `displayName` or name that you specified in the .json config file.
3. Enter values for any parameters that were configured in the config file.


Node properties
Transform 1
Output schema
Data preview
Data source - S3 bucket
Amazon S3
Transform - Dynamic Trans...
My Transform

Email Address

Enter your work email address below

Phone Number

Enter your mobile phone number below

Your age

Your gender

Your origin country ? - *optional*

What country were you born in?

Choose options

Do you want to receive promotional newsletter from us? - *optional*

☐

Usage examples

The following is an example of all possible parameters in a .json config file.

```
{
  "name": "MyTransform",
  "displayName": "My Transform",
  "description": "This transform description will be displayed in UI",
  "functionName": "myTransform",
  "parameters": [
      {
      "name": "email",
      "displayName": "Email Address",
      "type": "str",
      "description": "Enter your work email address below",
      "validationType": "RegularExpression",
      "validationRule": "^\\w+([\\.-]?\\w+)*@\\w+([\\.-]?\\w+)*(\\.\\w{2,3})+$",
      "validationMessage": "Please enter a valid email address"
    },
    {
      "name": "phone",
      "displayName": "Phone Number",
      "type": "str",
      "description": "Enter your mobile phone number below",
      "validationRule": "^\\(?(\\d{3})\\)?[- ]?(\\d{3})[- ]?(\\d{4})$",
      "validationMessage": "Please enter a valid US number"
```

```
     },
     {
       "name": "age",
       "displayName": "Your age",
       "type": "int",
       "isOptional": true
     },
     {
       "name": "gender",
       "displayName": "Your gender",
       "type": "str",
       "listOptions": [
             {"label": "Male", "value": "male"},
             {"label": "Female", "value": "female"},
             {"label": "Other", "value": "other"}
         ],
       "isOptional": true
     },
     {
       "name": "country",
       "displayName": "Your origin country ?",
       "type": "list",
       "listOptions": "Afghanistan,Albania,Algeria,American
  Samoa,Andorra,Angola,Anguilla,Antarctica,Antigua and
  Barbuda,Argentina,Armenia,Aruba,Australia,Austria,Azerbaijan,Bahamas,Bahrain,Bangladesh,Barbad
  and Herzegovina,Botswana,Bouvet Island,Brazil,British Indian Ocean Territory,Brunei
  Darussalam,Bulgaria,Burkina Faso,Burundi,Cambodia,Cameroon,Canada,Cape
  Verde,Cayman Islands,Central African Republic,Chad,Chile,China,Christmas
  Island,Cocos (Keeling Islands),Colombia,Comoros,Congo,Cook Islands,Costa
  Rica,Cote D'Ivoire (Ivory Coast),Croatia (Hrvatska,Cuba,Cyprus,Czech
  Republic,Denmark,Djibouti,Dominica,Dominican Republic,East Timor,Ecuador,Egypt,El
  Salvador,Equatorial Guinea,Eritrea,Estonia,Ethiopia,Falkland Islands (Malvinas),Faroe
  Islands,Fiji,Finland,France,France,Metropolitan,French Guiana,French Polynesia,French
  Southern
  Territories,Gabon,Gambia,Georgia,Germany,Ghana,Gibraltar,Greece,Greenland,Grenada,Guadeloupe,G
 Bissau,Guyana,Haiti,Heard and McDonald Islands,Honduras,Hong
  Kong,Hungary,Iceland,India,Indonesia,Iran,Iraq,Ireland,Israel,Italy,Jamaica,Japan,Jordan,Kazak
  (North),Korea
  (South),Kuwait,Kyrgyzstan,Laos,Latvia,Lebanon,Lesotho,Liberia,Libya,Liechtenstein,Lithuania,Lu
  Islands,Martinique,Mauritania,Mauritius,Mayotte,Mexico,Micronesia,Moldova,Monaco,Mongolia,Mont
  Antilles,New Caledonia,New Zealand,Nicaragua,Niger,Nigeria,Niue,Norfolk
  Island,Northern Mariana Islands,Norway,Oman,Pakistan,Palau,Panama,Papua
  New Guinea,Paraguay,Peru,Philippines,Pitcairn,Poland,Portugal,Puerto
  Rico,Qatar,Reunion,Romania,Russian Federation,Rwanda,Saint Kitts and Nevis,Saint
 Lucia,Saint Vincent and The Grenadines,Samoa,San Marino,Sao Tome and Principe,Saudi
 Arabia,Senegal,Seychelles,Sierra Leone,Singapore,Slovak Republic,Slovenia,Solomon
 Islands,Somalia,South Africa,S. Georgia and S. Sandwich Isls.,Spain,Sri
 Lanka,St. Helena,St. Pierre and Miquelon,Sudan,Suriname,Svalbard and Jan Mayen
 Islands,Swaziland,Sweden,Switzerland,Syria,Tajikistan,Tanzania,Thailand,Togo,Tokelau,Tonga,Tri
 and Tobago,Tunisia,Turkey,Turkmenistan,Turks and Caicos
 Islands,Tuvalu,Uganda,Ukraine,United Arab Emirates,United Kingdom
 (Britain / UK),United States of America (USA),US Minor Outlying
 Islands,Uruguay,Uzbekistan,Vanuatu,Vatican City State (Holy See),Venezuela,Viet
 Nam,Virgin Islands (British),Virgin Islands (US),Wallis and Futuna Islands,Western
 Sahara,Yemen,Yugoslavia,Zaire,Zambia,Zimbabwe",
      "description": "What country were you born in?",
      "listType": "str",
      "isOptional": true
    },
    {
      "name": "promotion",
      "displayName": "Do you want to receive promotional newsletter from us?",
      "type": "bool",
      "isOptional": true
    }
  ]
}
```

Examples of custom visual scripts

The following examples perform equivalent transformations. However, the second example (SparkSQL) is the cleanest and most efficient, followed by the Pandas UDF and finally the low level mapping in the first example. The following example is a complete example of a simple transformation to add up two columns:

```
from awsglue import DynamicFrame

# You can have other auxiliary variables, functions or classes on this file, it won't
 affect the runtime
def record_sum(rec, col1, col2, resultCol):
    rec[resultCol] = rec[col1] + rec[col2]
    return rec


# The number and name of arguments must match the definition on json config file
# (expect self which is the current DynamicFrame to transform
# If an argument is optional, you need to define a default value here
#  (resultCol in this example is an optional argument)
def custom_add_columns(self, col1, col2, resultCol="result"):
    # The mapping will alter the columns order, which could be important
    fields = [field.name for field in self.schema()]
    if resultCol not in fields:
        # If it's a new column put it at the end
        fields.append(resultCol)
    return self.map(lambda record: record_sum(record, col1, col2,
 resultCol)).select_fields(paths=fields)


# The name we assign on DynamicFrame must match the configured "functionName"
DynamicFrame.custom_add_columns = custom_add_columns
```

The following example is an equivalent transform leveraging the SparkSQL API.

```
from awsglue import DynamicFrame

# The number and name of arguments must match the definition on json config file
# (expect self which is the current DynamicFrame to transform
# If an argument is optional, you need to define a default value here
#  (resultCol in this example is an optional argument)
def custom_add_columns(self, col1, col2, resultCol="result"):
    df = self.toDF()
    return DynamicFrame.fromDF(
        df.withColumn(resultCol, df[col1] + df[col2]) # This is the conversion logic
        , self.glue_ctx, self.name)


# The name we assign on DynamicFrame must match the configured "functionName"
DynamicFrame.custom_add_columns = custom_add_columns
```

The following example uses the same transformations but using a pandas UDF, which is more efficient that using a plain UDF. For more information about writing pandas UDFs see: Apache Spark SQL documentation.

```
from awsglue import DynamicFrame
import pandas as pd
from pyspark.sql.functions import pandas_udf
```

```
# The number and name of arguments must match the definition on json config file
# (expect self which is the current DynamicFrame to transform
# If an argument is optional, you need to define a default value here
#  (resultCol in this example is an optional argument)
def custom_add_columns(self, col1, col2, resultCol="result"):
    @pandas_udf("integer")  # We need to declare the type of the result column
    def add_columns(value1: pd.Series, value2: pd.Series) # pd.Series:
        return value1 + value2

    df = self.toDF()
    return DynamicFrame.fromDF(
        df.withColumn(resultCol, add_columns(col1, col2)) # This is the conversion
 logic
        , self.glue_ctx, self.name)

# The name we assign on DynamicFrame must match the configured "functionName"
DynamicFrame.custom_add_columns = custom_add_columns
```

Video

The following video provides an introduction to visual custom transforms and demonstrates how to use them.

Using Data Lake frameworks with AWS Glue Studio

Overview

Open source data lake frameworks simplify incremental data processing for files stored in data lakes built on Amazon S3. AWS Glue 3.0 and later supports the following open-source data lake storage frameworks:

- Apache Hudi
- Linux Foundation Delta Lake
- Apache Iceberg

As of AWS Glue 4.0, AWS Glue provides native support for these frameworks so that you can read and write data that you store in Amazon S3 in a transactionally consistent manner. There's no need to install a separate connector or complete extra configuration steps in order to use these frameworks in AWS Glue jobs.

Data Lake frameworks can be used as a source or a target within AWS Glue Studio through Spark Script Editor jobs. For more information on using Apache Hudi, Apache Iceberg and Delta Lake see: Using data lake frameworks with AWS Glue ETL jobs.

Creating open table formats from an AWS Glue Streaming source

AWS Glue streaming ETL jobs continuously consume data from streaming sources, clean and transform the data in-flight, and make it available for analysis in seconds.

AWS offers a broad selection of services to support your needs. A database replication service such as AWS Database Migration Service can replicate the data from your source systems to Amazon S3, which commonly hosts the storage layer of the data lake. Although it’s straightforward to apply updates on a relational database management system (RDBMS) that backs an online source application, it's difficult to apply this CDC process on your data lakes. The open-source data management frameworks simplify incremental data processing and data pipeline development, and are a good option to solve this problem.

For more information, see:

- Create an Apache Hudi-based near-real-time transactional data lake using AWS Glue Streaming
- Build a real-time GDPR-aligned Apache Iceberg data lake

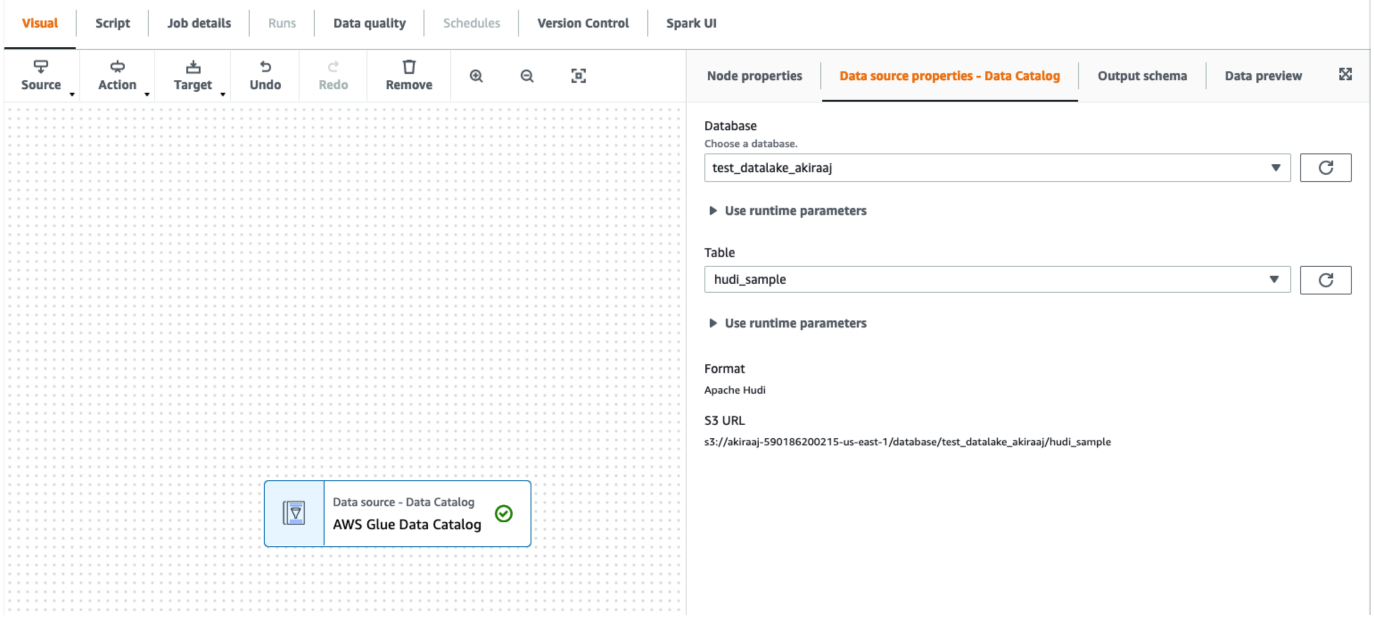
Using Hudi framework in AWS Glue Studio

When creating or editing a job, AWS Glue Studio automatically adds the corresponding Hudi libraries for you depending on the version of AWS Glue you are using. For more information, see Using the Hudi framework in AWS Glue.

Using Apache Hudi framework in Data Catalog data sources

To add a Hudi data source format to a job:

1. From the Source menu, choose AWS Glue Studio Data Catalog.
2. In the **Data source properties** tab, choose a database and table.
3. AWS Glue Studio displays the format type as Apache Hudi and the Amazon S3 URL.


Visual
Script
Job details
Runs
Data quality
Schedules
Version Control
Spark UI
Source
Action
Target
Undo
Redo
Remove
Node properties
Data source properties - Data Catalog
Output schema
Data preview
Use runtime parameters
Use runtime parameters
Data source - Data Catalog
AWS Glue Data Catalog

Database
Choose a database.
test_datalake_akiraaj

Table
hudi_sample

Format
Apache Hudi

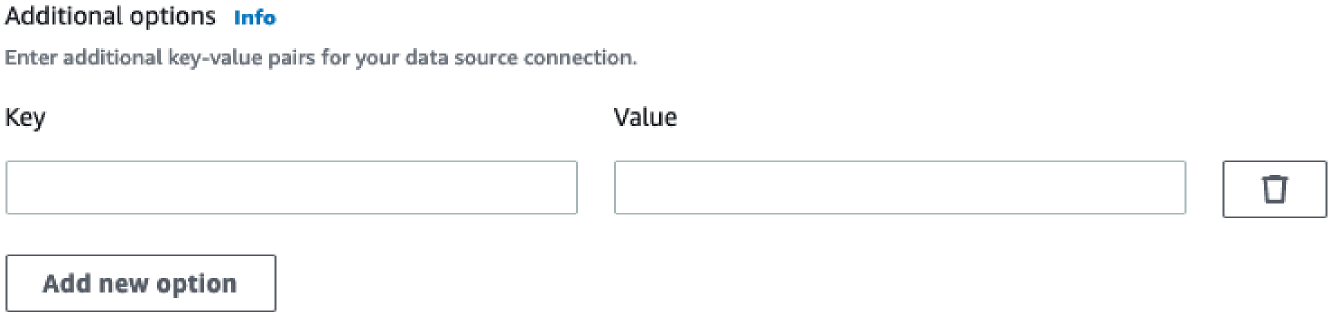
S3 URL
s3://akiraaj-590186200215-us-east-1/database/test_datalake_akiraaj/hudi_sample

Using Hudi framework in Amazon S3 data sources

1. From the Source menu, choose Amazon S3.
2. If you choose Data Catalog table as the Amazon S3 source type, choose a database and table.
3. AWS Glue Studio displays the format as Apache Hudi and the Amazon S3 URL.
4. If you choose Amazon S3 location as the **Amazon S3 source type**, choose the Amazon S3 URL by clicking **Browse Amazon S3**.
5. In **Data format**, select Apache Hudi.

> **Note**
>
> If AWS Glue Studio is unable to infer the schema from the Amazon S3 folder or file you selected, choose **Additional options** to select a new folder or file.
> In **Additional options** choose from the following options under **Schema inference**:
>
> - Let AWS Glue Studio automatically choose a sample file — AWS Glue Studio will choose a sample file in the Amazon S3 location so that the schema can be inferred. In the **Auto-sampled file** field, you can view the file that was automatically selected.
> - Choose a sample file from Amazon S3 — choose the Amazon S3 file to use by clicking **Browse Amazon S3**.

6. Click **Infer schema**. You can then view the output schema by clicking on the **Output schema** tab.
7. Choose **Additional options** to enter a key-value pair.



Additional options Info

Enter additional key-value pairs for your data source connection.

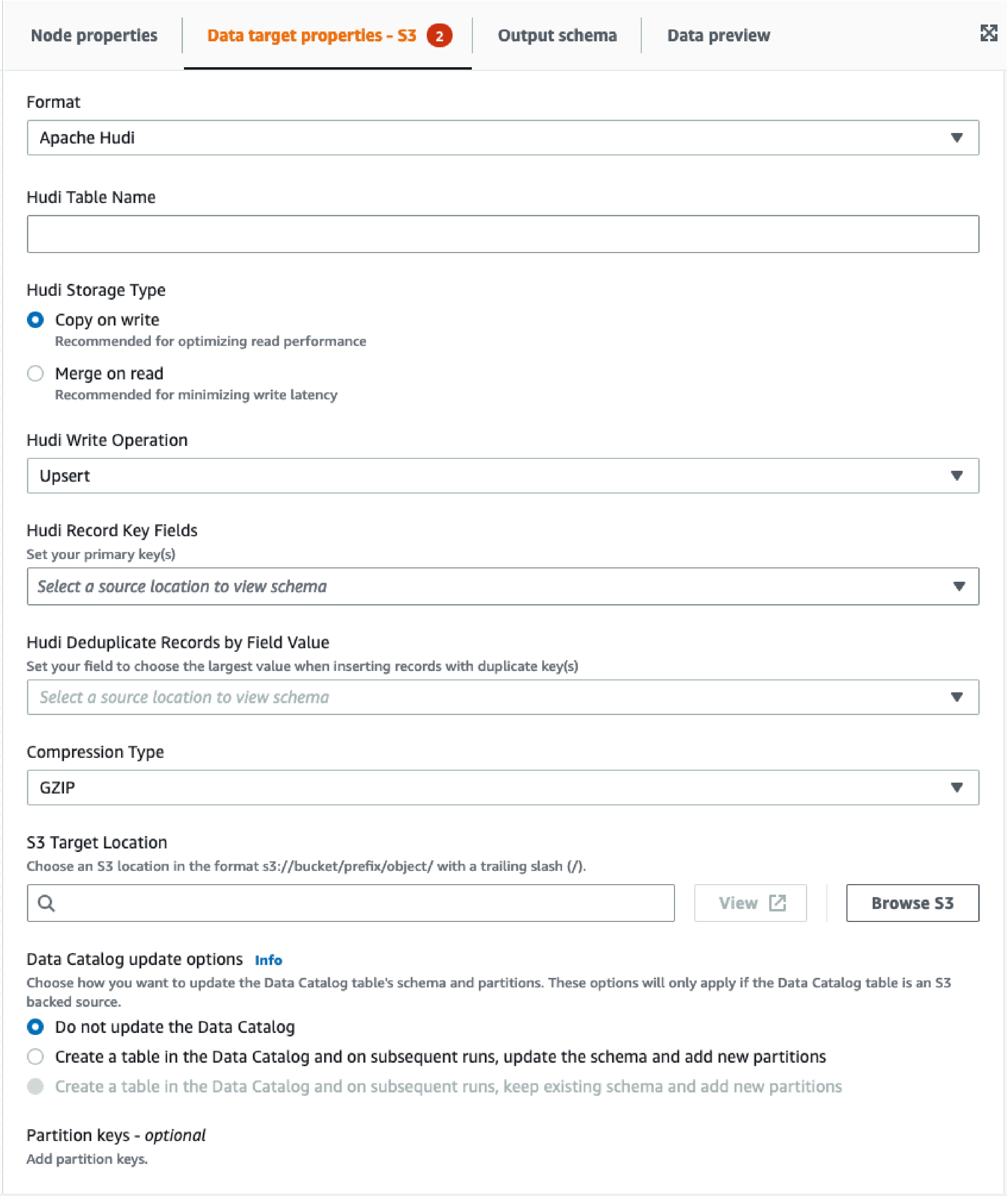
Using Apache Hudi framework in data targets

Using Apache Hudi framework in Data Catalog data targets

1. From the **Target** menu, choose AWS Glue Studio Data Catalog.
2. In the **Data source properties** tab, choose a database and table.
3. AWS Glue Studio displays the format type as Apache Hudi and the Amazon S3 URL.

Using Apache Hudi framework in Amazon S3 data targets

Enter values or select from the available options to configure Apache Hudi format. For more information on Apache Hudi, see [Apache Hudi documentation](#).



Node properties | **Data target properties - S3** 2 | Output schema | Data preview

Format

Apache Hudi ▼

Hudi Table Name

Hudi Storage Type

● Copy on write
Recommended for optimizing read performance

○ Merge on read
Recommended for minimizing write latency

Hudi Write Operation

Upsert ▼

Hudi Record Key Fields
Set your primary key(s)

Select a source location to view schema ▼

Hudi Deduplicate Records by Field Value
Set your field to choose the largest value when inserting records with duplicate key(s)

Select a source location to view schema ▼

Compression Type

GZIP ▼

S3 Target Location
Choose an S3 location in the format s3://bucket/prefix/object/ with a trailing slash (/).

View | **Browse S3**

Data Catalog update options **Info**
Choose how you want to update the Data Catalog table's schema and partitions. These options will only apply if the Data Catalog table is an S3 backed source.

● Do not update the Data Catalog
○ Create a table in the Data Catalog and on subsequent runs, update the schema and add new partitions
○ Create a table in the Data Catalog and on subsequent runs, keep existing schema and add new partitions

Partition keys - *optional*
Add partition keys.

- **Hudi Table Name** — this is the name of your hudi table.
- **Hudi Storage Type** — choose from two options:
    - **Copy on write** — recommended for optimizing read performance. This is the default Hudi storage type. Each update creates a new version of files during a write.
    - **Merge on read** — recommended for minimizing write latency. Updates are logged to row-based delta files and are compacted as needed to create new versions of the columnar files.
- **Hudi Write Operation** - choose from the following options:
    - **Upsert** — this is the default operation where the input records are first tagged as inserts or updates by looking up the index. Recommended where you are updating existing data.
    - **Insert** — this inserts records but doesn't check for existing records and may result in duplicates.
    - **Bulk Insert** — this inserts records and is recommended for large amounts of data.
- **Hudi Record Key Fields** — use the search bar to search for and choose primary record keys. Records in Hudi are identified by a primary key which is a pair of record key and partition path where the record belongs to.
- **Hudi Precombine Field** — this is the field used in preCombining before actual write. When two records have the same key value, AWS Glue Studio will pick the one with the largest value for the precombine field. Set a field with incremental value (e.g. updated_at) belongs to.
- **Compression Type** — choose from one of the compression type options: Uncompressed, GZIP, LZO, or Snappy.
- **Amazon S3 Target Location** — choose the Amazon S3 target location by clicking **Browse S3**.
- **Data Catalog update options** — choose from the following options:
    - Do not update the Data Catalog: (Default) Choose this option if you don't want the job to update the Data Catalog, even if the schema changes or new partitions are added.
    - Create a table in the Data Catalog and on subsequent runs, update the schema and add new partitions: I f you choose this option, the job creates the table in the Data Catalog on the first run of the job. On subsequent job runs, the job updates the Data Catalog table if the schema changes or new partitions are added.

      You must also select a database from the Data Catalog and enter a table name.
    - Create a table in the Data Catalog and on subsequent runs, keep existing schema and add new partitions: If you choose this option, the job creates the table in the Data Catalog on the first run of the job. On subsequent job runs, the job updates the Data Catalog table only to add new partitions.

You must also select a database from the Data Catalog and enter a table name.

- Partition keys: Choose which columns to use as partitioning keys in the output. To add more partition keys, choose Add a partition key.
- **Addtional options** — enter a key-value pair as needed.

Generating code through AWS Glue Studio

When the job is saved, the following job parameters are added to the job if a Hudi source or target are detected:

- `--datalake-formats` – a distinct list of data lake formats detected in the visual job (either directly by choosing a “Format” or indirectly by selecting a catalog table that is backed by a data lake).
- `--conf` – generated based on the value of `--datalake-formats`. For example, if the value for `--datalake-formats` is 'hudi', AWS Glue generates a value of `spark.serializer=org.apache.spark.serializer.KryoSerializer —conf spark.sql.hive.convertMetastoreParquet=false` for this parameter.

Overriding AWS Glue-provided libraries

To use a version of Hudi that AWS Glue doesn't support, you can specify your own Hudi library JAR files. To use your own JAR file:

- use the `--extra-jars` job parameter. For example, `'--extra-jars': 's3pathtojarfile.jar'`. For more information, see AWS Glue job parameters.
- do not include `hudi` as a value for the `--datalake-formats` job parameter. Entering a blank string as a value ensures that no data lake libraries are provided for you by AWS Glue automatically. For more information, see Using the Hudi framework in AWS Glue.

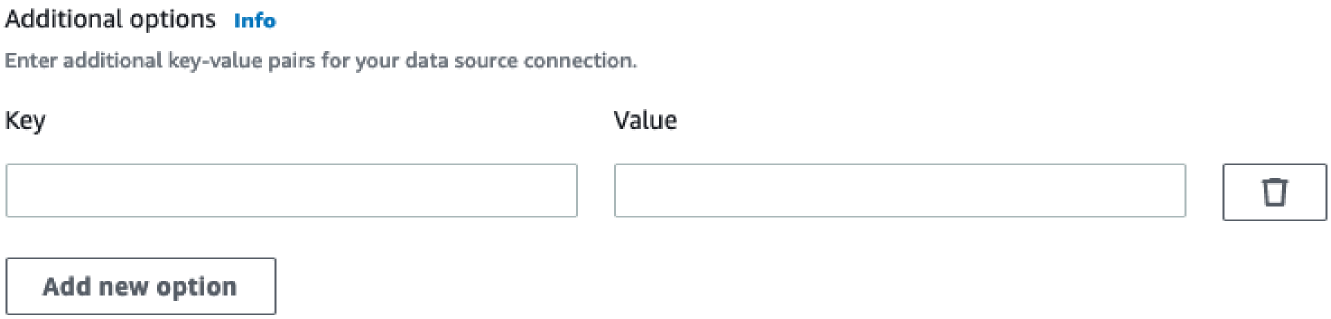
Using Delta Lake framework in AWS Glue Studio

Using Delta Lake framework in data sources

Using Delta Lake framework in Amazon S3 data sources

1. From the Source menu, choose Amazon S3.

2. If you choose Data Catalog table as the Amazon S3 source type, choose a database and table.
3. AWS Glue Studio displays the format as Delta Lake and the Amazon S3 URL.
4. Choose **Additional options** to enter a key-value pair. For example, a key-value pair could be: **key**: timestampAsOf and **value**: 2023-02-24 14:16:18.



Additional options Info

Enter additional key-value pairs for your data source connection.

5. If you choose Amazon S3 location as the **Amazon S3 source type**, choose the Amazon S3 URL by clicking **Browse Amazon S3**.
6. In **Data format**, choose Delta Lake.

> **Note**
>
> If AWS Glue Studio is unable to infer the schema from the Amazon S3 folder or file you selected, choose **Additional options** to select a new folder or file.
> In **Additional options** choose from the following options under **Schema inference**:
>
> - Let AWS Glue Studio automatically choose a sample file — AWS Glue Studio will choose a sample file in the Amazon S3 location so that the schema can be inferred. In the **Auto-sampled file** field, you can view the file that was automatically selected.
> - Choose a sample file from Amazon S3 — choose the Amazon S3 file to use by clicking **Browse Amazon S3**.

7. Click **Infer schema**. You can then view the output schema by clicking on the **Output schema** tab.

Using Delta Lake framework in Data Catalog data sources

1. From the **Source** menu, choose AWS Glue Studio Data Catalog.
2. In the **Data source properties** tab, choose a database and table.

3. AWS Glue Studio displays the format type as Delta Lake and the Amazon S3 URL.

> **Note**
>
> If your Delta Lake source is not registered as the AWS Glue Data Catalog table yet, you have two options:
>
> 1. Create a AWS Glue crawler for the Delta Lake data store. For more information, see How to specify configuration options for a Delta Lake data store.
> 2. Use an Amazon S3 data source to select your Delta Lake data source. See Using Delta Lake framework in Amazon S3 data sources .

Using Delta Lake formats in data targets

Using Delta Lake formats in Data Catalog data targets

1. From the **Target** menu, choose AWS Glue Studio Data Catalog.
2. In the **Data source properties** tab, choose a database and table.
3. AWS Glue Studio displays the format type as Delta Lake and the Amazon S3 URL.

Using Delta Lake formats in Amazon S3 data sources

Enter values or select from the available options to configure Delta Lake format.

- **Compression Type** — choose from one of the compression type options: Uncompressed or Snappy.
- **Amazon S3 Target Location** — choose the Amazon S3 target location by clicking **Browse S3**.
- **Data Catalog update options** — updating the Data Catalog is not supported for this format in the Glue Studio visual editor.
  - Do not update the Data Catalog: (Default) Choose this option if you don't want the job to update the Data Catalog, even if the schema changes or new partitions are added.
  - To update the Data Catalog after the AWS Glue job execution, run or schedule a AWS Glue crawler. For more information, see How to specify configuration options for a Delta Lake data store.
- **Partition keys** — Choose which columns to use as partitioning keys in the output. To add more partition keys, choose **Add a partition key**.

- Optionally, choose **Addtional options** to enter a key-value pair. For example, a key-value pair could be: **key**: timestampAsOf and **value**: 2023-02-24 14:16:18.

Using Apache Iceberg framework in AWS Glue Studio

Using Apache Iceberg framework in data targets

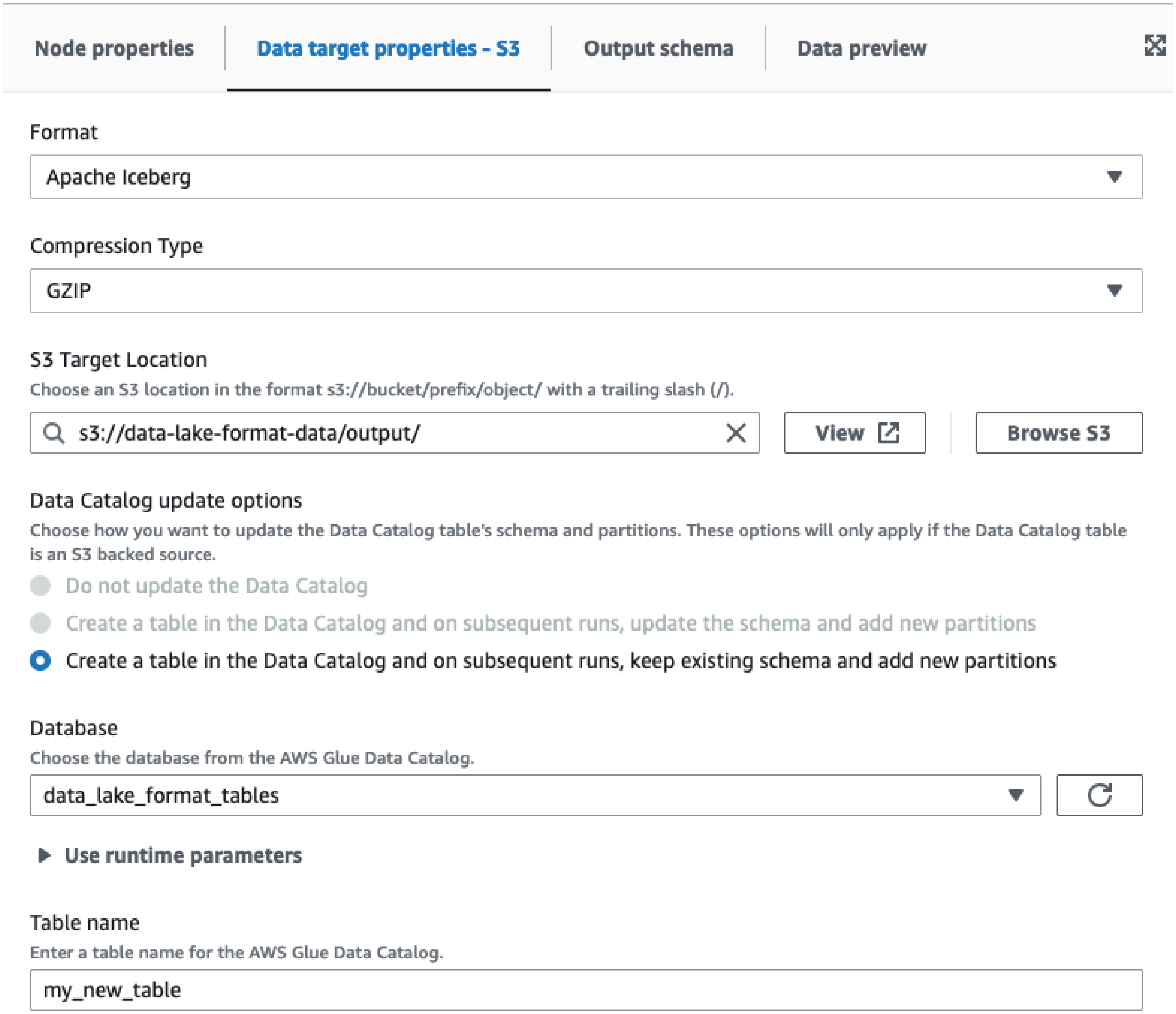
Using Apache Iceberg framework in Data Catalog data targets

1. From the **Target** menu, choose AWS Glue Studio Data Catalog.
2. In the **Data source properties** tab, choose a database and table.
3. AWS Glue Studio displays the format type as Apache Iceberg and the Amazon S3 URL.

Using Apache Iceberg framework in Amazon S3 data targets

Enter values or select from the available options to configure Apache Iceberg format.

- **Format –** choose **Apache Iceberg** from the drop-down menu.
- **Amazon S3 Target Location –** choose the Amazon S3 target location by clicking **Browse S3**.
- **Data Catalog update options – Create a table in the Data Catalog and on subsequent runs, keep existing schema and add new partitions** must be selected to proceed. Writing a new Iceberg table using AWS Glue requires the Data Catalog to be configured as the catalog for the Iceberg table. To update an existing Iceberg table that has been registered in the Data Catalog, choose Data Catalog as the target.
  - **Database** – Choose the database from the Data Catalog.
  - **Table Name –** Enter the value for your table name. Apache Iceberg table names must be in all lower case. Use underscores if needed since spaces are not allowed. For example "data_lake_format_tables".



Format

Apache Iceberg

Compression Type

GZIP

S3 Target Location

Choose an S3 location in the format s3://bucket/prefix/object/ with a trailing slash (/).

s3://data-lake-format-data/output/

Data Catalog update options

Choose how you want to update the Data Catalog table's schema and partitions. These options will only apply if the Data Catalog table is an S3 backed source.

○ Do not update the Data Catalog

○ Create a table in the Data Catalog and on subsequent runs, update the schema and add new partitions

● Create a table in the Data Catalog and on subsequent runs, keep existing schema and add new partitions

Database

Choose the database from the AWS Glue Data Catalog.

data_lake_format_tables

Table name

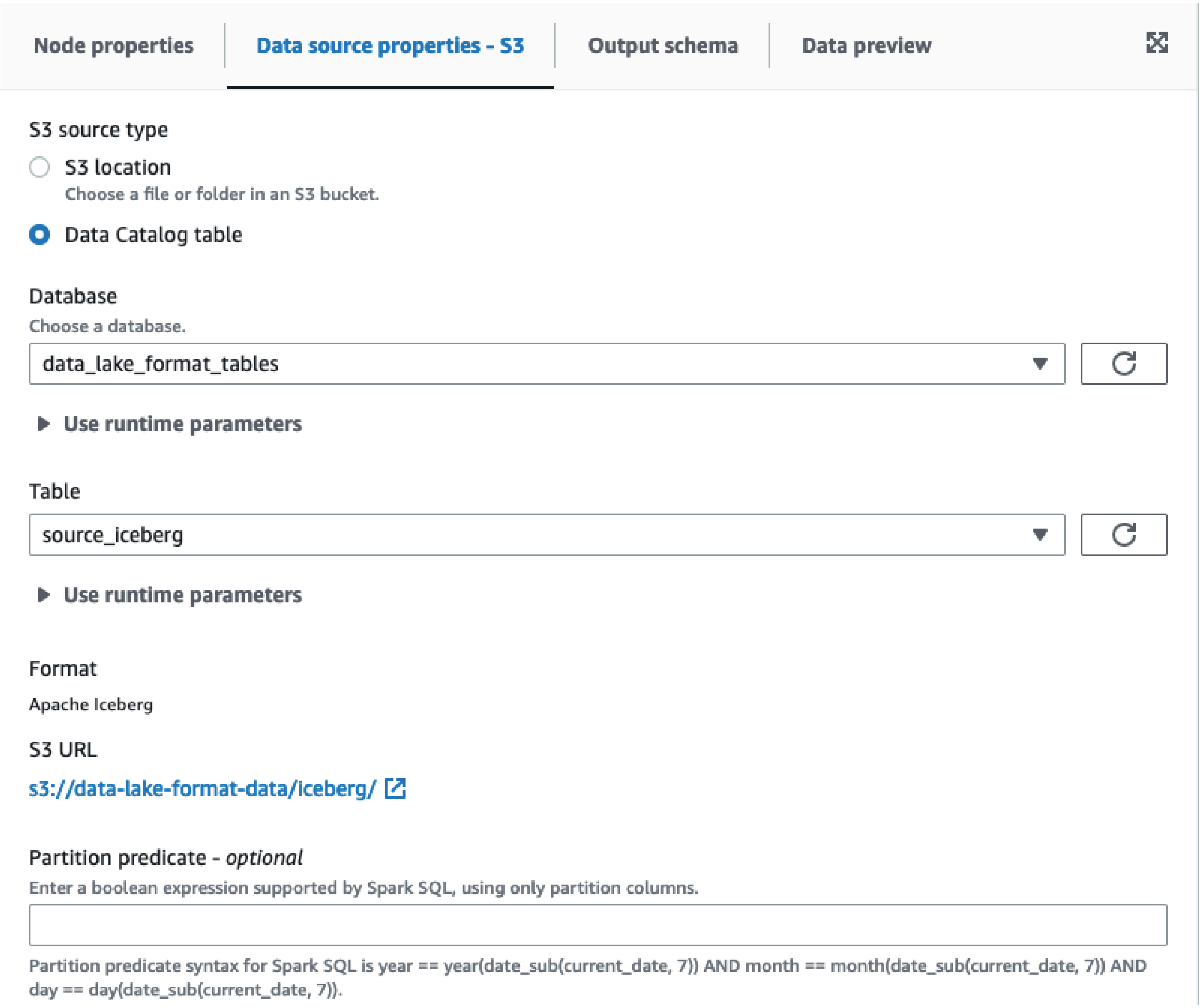
Enter a table name for the AWS Glue Data Catalog.

my_new_table

Using Apache Iceberg framework in Amazon S3 data sources

Using Apache Iceberg framework in Data Catalog data sources

1. From the **Source** menu, choose AWS Glue Studio Data Catalog.
2. In the **Data source properties** tab, choose a database and table.
3. AWS Glue Studio displays the format type as Apache Iceberg and the Amazon S3 URL.


Node properties
Data source properties - S3
Output schema
Data preview

S3 source type

○ S3 location
Choose a file or folder in an S3 bucket.

● Data Catalog table

Database
Choose a database.
data_lake_format_tables

▶ Use runtime parameters

Table
source_iceberg

▶ Use runtime parameters

Format
Apache Iceberg

S3 URL
s3://data-lake-format-data/iceberg/

Partition predicate - *optional*
Enter a boolean expression supported by Spark SQL, using only partition columns.

Partition predicate syntax for Spark SQL is year == year(date_sub(current_date, 7)) AND month == month(date_sub(current_date, 7)) AND day == day(date_sub(current_date, 7)).

Using Apache Iceberg framework in Amazon S3 data sources

Apache Iceberg is not available as a data option for Amazon S3 source nodes in AWS Glue Studio.

Connecting to data sources using Visual ETL jobs

While creating a new job, you can use connections to connect to data when editing visual ETL jobs in AWS Glue. You can do this by adding source nodes that use connectors to read in data, and target nodes to specify the location for writing out data.

Topics

- [Modifying properties of a data source node]

- Using Data Catalog tables for the data source
- Using a connector for the data source
- Using files in Amazon S3 for the data source
- Using a streaming data source
- References

Modifying properties of a data source node

To specify the data source properties, you first choose a data source node in the job diagram. Then, on the right side in the node details panel, you configure the node properties.

To modify the properties of a data source node

1. Go to the visual editor for a new or saved job.
2. Choose a data source node in the job diagram.
3. Choose the **Node properties** tab in the node details panel, and then enter the following information:
   - **Name**: (Optional) Enter a name to associate with the node in the job diagram. This name should be unique among all the nodes for this job.
   - **Node type**: The node type determines the action that is performed by the node. In the list of options for **Node type**, choose one of the values listed under the heading **Data source**.
4. Configure the **Data source properties** information. For more information, see the following sections:
   - Using Data Catalog tables for the data source
   - Using a connector for the data source
   - Using files in Amazon S3 for the data source
   - Using a streaming data source
5. (Optional) After configuring the node properties and data source properties, you can view the schema for your data source by choosing the **Output schema** tab in the node details panel. The first time you choose this tab for any node in your job, you are prompted to provide an IAM role to access the data. If you have not specified an IAM role on the **Job details** tab, you are prompted to enter an IAM role here.

6. (Optional) After configuring the node properties and data source properties, you can preview the dataset from your data source by choosing the **Data preview** tab in the node details panel. The first time you choose this tab for any node in your job, you are prompted to provide an IAM role to access the data. There is a cost associated with using this feature, and billing starts as soon as you provide an IAM role.

Using Data Catalog tables for the data source

For all data sources except Amazon S3 and connectors, a table must exist in the AWS Glue Data Catalog for the source type that you choose. AWS Glue does not create the Data Catalog table.

To configure a data source node based on a Data Catalog table

1. Go to the visual editor for a new or saved job.
2. Choose a data source node in the job diagram.
3. Choose the **Data source properties** tab, and then enter the following information:

    - **S3 source type**: (For Amazon S3 data sources only) Choose the option **Select a Catalog table** to use an existing AWS Glue Data Catalog table.
    - **Database**: Choose the database in the Data Catalog that contains the source table you want to use for this job. You can use the search field to search for a database by its name.
    - **Table**: Choose the table associated with the source data from the list. This table must already exist in theAWS Glue Data Catalog. You can use the search field to search for a table by its name.
    - **Partition predicate**: (For Amazon S3 data sources only) Enter a Boolean expression based on Spark SQL that includes only the partitioning columns. For example: `"(year=='2020' and month=='04')"`
    - **Temporary directory**: (For Amazon Redshift data sources only) Enter a path for the location of a working directory in Amazon S3 where your ETL job can write temporary intermediate results.
    - **Role associated with the cluster**: (For Amazon Redshift data sources only) Enter a role for your ETL job to use that contains permissions for Amazon Redshift clusters. For more information, see the section called “Data source and data target permissions”.

Using a connector for the data source

If you select a connector for the **Node type**, follow the instructions at Authoring jobs with custom connectors to finish configuring the data source properties.

Using files in Amazon S3 for the data source

If you choose Amazon S3 as your data source, then you can choose either:

- A Data Catalog database and table.
- A bucket, folder, or file in Amazon S3.

If you use an Amazon S3 bucket as your data source, AWS Glue detects the schema of the data at the specified location from one of the files, or by using the file you specify as a sample file. Schema detection occurs when you use the **Infer schema** button. If you change the Amazon S3 location or the sample file, then you must choose **Infer schema** again to perform the schema detection using the new information.

To configure a data source node that reads directly from files in Amazon S3

1. Go to the visual editor for a new or saved job.
2. Choose a data source node in the job diagram for an Amazon S3 source.
3. Choose the **Data source properties** tab, and then enter the following information:
   - **S3 source type**: (For Amazon S3 data sources only) Choose the option **S3 location**.
   - **S3 URL**: Enter the path to the Amazon S3 bucket, folder, or file that contains the data for your job. You can choose **Browse S3** to select the path from the locations available to your account.
   - **Recursive**: Choose this option if you want AWS Glue to read data from files in child folders at the S3 location.

     If the child folders contain partitioned data, AWS Glue doesn't add any partition information that's specified in the folder names to the Data Catalog. For example, consider the following folders in Amazon S3:

     ```
     S3://sales/year=2019/month=Jan/day=1
     S3://sales/year=2019/month=Jan/day=2
     ```

If you choose **Recursive** and select the `sales` folder as your S3 location, then AWS Glue reads the data in all the child folders, but doesn't create partitions for year, month or day.

- **Data format**: Choose the format that the data is stored in. You can choose JSON, CSV, or Parquet. The value you select tells the AWS Glue job how to read the data from the source file.

  > **Note**
  >
  > If you don't select the correct format for your data, AWS Glue might infer the schema correctly, but the job won't be able to correctly parse the data from the source file.

  You can enter additional configuration options, depending on the format you choose.

  - **JSON** (JavaScript Object Notation)
    - **JsonPath**: Enter a JSON path that points to an object that is used to define a table schema. JSON path expressions always refer to a JSON structure in the same way as XPath expression are used in combination with an XML document. The "root member object" in the JSON path is always referred to as $, even if it's an object or array. The JSON path can be written in dot notation or bracket notation.

      For more information about the JSON path, see JsonPath on the GitHub website.
    - **Records in source files can span multiple lines**: Choose this option if a single record can span multiple lines in the CSV file.
  - **CSV** (comma-separated values)
    - **Delimiter**: Enter a character to denote what separates each column entry in the row, for example, `;` or `,`.
    - **Escape character**: Enter a character that is used as an escape character. This character indicates that the character that immediately follows the escape character should be taken literally, and should not be interpreted as a delimiter.
    - **Quote character**: Enter the character that is used to group separate strings into a single value. For example, you would choose **Double quote (")** if you have values such as `"This is a single value"` in your CSV file.
    - **Records in source files can span multiple lines**: Choose this option if a single record can span multiple lines in the CSV file.

    - **First line of source file contains column headers**: Choose this option if the first row in the CSV file contains column headers instead of data.
  - **Parquet** (Apache Parquet columnar storage)

    There are no additional settings to configure for data stored in Parquet format.
  - **Apache Hudi**

    There are no additional settings to configure for data stored in Apache Hudi format.
  - **Delta Lake**

    There are no additional settings to configure for data stored in Delta Lake format.
  - **Excel**

    There are no additional settings to configure for data stored in Excel format.
- **Partition predicate**: To partition the data that is read from the data source, enter a Boolean expression based on Spark SQL that includes only the partitioning columns. For example: `"(year=='2020' and month=='04')"`
- **Advanced options**: Expand this section if you want AWS Glue to detect the schema of your data based on a specific file.
  - **Schema inference**: Choose the option **Choose a sample file from S3** if you want to use a specific file instead of letting AWS Glue choose a file. Schema inference is not available for the Excel source.
  - **Auto-sampled file**: Enter the path to the file in Amazon S3 to use for inferring the schema.

  If you're editing a data source node and change the selected sample file, choose **Reload schema** to detect the schema by using the new sample file.

4. Choose the **Infer schema** button to detect the schema from the sources files in Amazon S3. If you change the Amazon S3 location or the sample file, you must choose **Infer schema** again to infer the schema using the new information.

Using a streaming data source

You can create streaming extract, transform, and load (ETL) jobs that run continuously and consume data from streaming sources in Amazon Kinesis Data Streams, Apache Kafka, and Amazon Managed Streaming for Apache Kafka (Amazon MSK).

To configure properties for a streaming data source

1. Go to the visual graph editor for a new or saved job.
2. Choose a data source node in the graph for Kafka or Kinesis Data Streams.
3. Choose the **Data source properties** tab, and then enter the following information:

   Kinesis

   - **Kinesis source type**: Choose the option **Stream details** to use direct access to the streaming source or choose **Data Catalog table** to use the information stored there instead.

     If you choose **Stream details**, specify the following additional information.

     - **Location of data stream**: Choose whether the stream is associated with the current user, or if it is associated with a different user.
     - **Region**: Choose the AWS Region where the stream exists. This information is used to construct the ARN for accessing the data stream.
     - **Stream ARN**: Enter the Amazon Resource Name (ARN) for the Kinesis data stream. If the stream is located within the current account, you can choose the stream name from the drop-down list. You can use the search field to search for a data stream by its name or ARN.
     - **Data format**: Choose the format used by the data stream from the list.

       AWS Glue automatically detects the schema from the streaming data.

     If you choose **Data Catalog table**, specify the following additional information.

     - **Database**: (Optional) Choose the database in the AWS Glue Data Catalog that contains the table associated with your streaming data source. You can use the search field to search for a database by its name.
     - **Table**: (Optional) Choose the table associated with the source data from the list. This table must already exist in the AWS Glue Data Catalog. You can use the search field to search for a table by its name.
     - **Detect schema**: Choose this option to have AWS Glue detect the schema from the streaming data, rather than using the schema information in a Data Catalog table. This option is enabled automatically if you choose the **Stream details** option.

- **Starting position**: By default, the ETL job uses the **Earliest** option, which means it reads data starting with the oldest available record in the stream. You can instead choose **Latest**, which indicates the ETL job should start reading from just after the most recent record in the stream.
- **Window size**: By default, your ETL job processes and writes out data in 100-second windows. This allows data to be processed efficiently and permits aggregations to be performed on data that arrives later than expected. You can modify this window size to increase timeliness or aggregation accuracy.

  AWS Glue streaming jobs use checkpoints rather than job bookmarks to track the data that has been read.
- **Connection options**: Expand this section to add key-value pairs to specify additional connection options. For information about what options you can specify here, see "connectionType": "kinesis" in the *AWS Glue Developer Guide*.

Kafka

- **Apache Kafka source**: Choose the option **Stream details** to use direct access to the streaming source or choose **Data Catalog table** to use the information stored there instead.

  If you choose **Data Catalog table**, specify the following additional information.
  - **Database**: (Optional) Choose the database in the AWS Glue Data Catalog that contains the table associated with your streaming data source. You can use the search field to search for a database by its name.
  - **Table**: (Optional) Choose the table associated with the source data from the list. This table must already exist in the AWS Glue Data Catalog. You can use the search field to search for a table by its name.
  - **Detect schema**: Choose this option to have AWS Glue detect the schema from the streaming data, rather than storing the schema information in a Data Catalog table. This option is enabled automatically if you choose the **Stream details** option.

  If you choose **Stream details**, specify the following additional information.
  - **Connection name**: Choose the AWS Glue connection that contains the access and authentication information for the Kafka data stream. You must use a connection with

    Kafka streaming data sources. If a connection doesn't exist, you can use the AWS Glue console to create a connection for your Kafka data stream.

  - **Topic name**: Enter the name of the topic to read from.
  - **Data format**: Choose the format to use when reading data from the Kafka event stream.
- **Starting position**: By default, the ETL job uses the **Earliest** option, which means it reads data starting with the oldest available record in the stream. You can instead choose **Latest**, which indicates the ETL job should start reading from just after the most recent record in the stream.
- **Window size**: By default, your ETL job processes and writes out data in 100-second windows. This allows data to be processed efficiently and permits aggregations to be performed on data that arrives later than expected. You can modify this window size to increase timeliness or aggregation accuracy.

  AWS Glue streaming jobs use checkpoints rather than job bookmarks to track the data that has been read.
- **Connection options**: Expand this section to add key-value pairs to specify additional connection options. For information about what options you can specify here, see "connectionType": "kafka" in the *AWS Glue Developer Guide*.

> **Note**
>
> Data previews are not currently supported for streaming data sources.

References

Best Practices

- Build an ETL service pipeline to load data incrementally from Amazon S3 to Amazon Redshift using AWS Glue

ETL programming

- Connection types and options for ETL in AWS Glue
- JDBC connectionType values

- Advanced options for moving data to and from Amazon Redshift

Configuring data target nodes

The data target is where the job writes the transformed data.

Overview of data target options

Your data target (also called a *data sink*) can be:

- **S3 –** The job writes the data in a file in the Amazon S3 location you choose and in the format you specify.

  If you configure partition columns for the data target, then the job writes the dataset to Amazon S3 into directories based on the partition key.
- **AWS Glue Data Catalog –** The job uses the information associated with the table in the Data Catalog to write the output data to a target location.

  You can create the table manually or with the crawler. You can also use AWS CloudFormation templates to create tables in the Data Catalog.
- A connector – A connector is a piece of code that facilitates communication between your data store and AWS Glue. The job uses the connector and associated connection to write the output data to a target location. You can either subscribe to a connector offered in AWS Marketplace, or you can create your own custom connector. For more information, see Adding connectors to AWS Glue Studio

You can choose to update the Data Catalog when your job writes to an Amazon S3 data target. Instead of requiring a crawler to update the Data Catalog when the schema or partitions change, this option makes it easy to keep your tables up to date. This option simplifies the process of making your data available for analytics by optionally adding new tables to the Data Catalog, updating table partitions, and updating the schema of your tables directly from the job.

Editing the data target node

The data target is where the job writes the transformed data.

To add or configure a data target node in your job diagram

1. (Optional) If you need to add a target node, choose **Target** in the toolbar at the top of the visual editor, and then choose either **S3** or **Glue Data Catalog**.

    - If you choose **S3** for the target, then the job writes the dataset to one or more files in the Amazon S3 location you specify.
    - If you choose **AWS Glue Data Catalog** for the target, then the job writes to a location described by the table selected from the Data Catalog.

2. Choose a data target node in the job diagram. When you choose a node, the node details panel appears on the right-side of the page.
3. Choose the **Node properties** tab, and then enter the following information:

    - **Name**: Enter a name to associate with the node in the job diagram.
    - **Node type**: A value should already be selected, but you can change it as needed.
    - **Node parents**: The parent node is the node in the job diagram that provides the output data you want to write to the target location. For a pre-populated job diagram, the target node should already have the parent node selected. If there is no parent node displayed, then choose a parent node from the list.

      A target node has a single parent node.

4. Configure the **Data target properties** information. For more information, see the following sections:

    - Using Amazon S3 for the data target
    - Using Data Catalog tables for the data target
    - Using a connector for the data target

5. (Optional) After configuring the data target node properties, you can view the output schema for your data by choosing the **Output schema** tab in the node details panel. The first time you choose this tab for any node in your job, you are prompted to provide an IAM role to access the data. If you have not specified an IAM role on the **Job details** tab, you are prompted to enter an IAM role here.

Using Amazon S3 for the data target

For all data sources except Amazon S3 and connectors, a table must exist in the AWS Glue Data Catalog for the source type that you choose. AWS Glue Studio does not create the Data Catalog table.

To configure a data target node that writes to Amazon S3

1. Go to the visual editor for a new or saved job.
2. Choose a data source node in the job diagram.
3. Choose the **Data source properties** tab, and then enter the following information:

   - **Format**: Choose a format from the list. The available format types for the data results are:
     - **JSON**: JavaScript Object Notation.
     - **CSV**: Comma-separated values.
     - **Avro**: Apache Avro JSON binary.
     - **Parquet**: A custom Parquet writer type that is optimized for `DynamicFrames` as the data format. Instead of requiring a precomputed schema for the data, it computes and modifies the schema dynamically.
     - **ORC**: Apache Optimized Row Columnar (ORC) format.
     - **Apache Hudi**: An open-source data lake storage framework that simplifies incremental data processing and data pipeline development.
     - **Apache Iceberg**: A high-performance table format that works just like an SQL table.
     - **Delta Lake**: An open-source data lake storage framework that helps you perform ACID transactions, scale metadata handling, and unify streaming and batch data processing.
     - **XML**: Extensible Markup Language (XML).
     - **Tableau Hyper**: Tableau's in-memory data engine technology.

     To learn more about these format options, see Format Options for ETL Inputs and Outputs in AWS Glue in the *AWS Glue Developer Guide*.

   - **Compression Type**: You can choose to optionally compress the data using the file types `CSV`, `JSON`, or `Parquet`. The default is no compression, or **None**.

File Type	Compressions
JSON/CSV/XML	GZIP, BZIP2, BROTLI, DEFLATE, LZ4, Snappy
Parquet	Snappy, LZ4, LZO, BROTLI, GZIP
ORC	Snappy, ZLIB, Uncompressed, LZO
Avro	GZIP, BZIP2, BROTLI, DEFLATE, LZ4, Snappy
Delta Lake	GZIP, BZIP2, BROTLI, DEFLATE, LZ4, Snappy
Apache Hudi	GZIP, LZO, Snappy
Apache Iceberg	GZIP, LZO, Snappy
Tableau Hyper	None

- **S3 Target Location**: The Amazon S3 bucket and location for the data output. You can choose the **Browse S3** button to see the Amazon S3 buckets that you have access to and choose one as the target destination.
- **Data catalog update options**
  - **Do not update the Data Catalog**: (Default) Choose this option if you don't want the job to update the Data Catalog, even if the schema changes or new partitions are added.
  - **Create a table in the Data Catalog and on subsequent runs, update the schema and add new partitions**: If you choose this option, the job creates the table in the Data Catalog on the first run of the job. On subsequent job runs, the job updates the Data Catalog table if the schema changes or new partitions are added.

    You must also select a database from the Data Catalog and enter a table name.
  - **Create a table in the Data Catalog and on subsequent runs, keep existing schema and add new partitions**: If you choose this option, the job creates the table in the Data Catalog on the first run of the job. On subsequent job runs, the job updates the Data Catalog table only to add new partitions.

    You must also select a database from the Data Catalog and enter a table name.
- **File Partitioning**: Choose which type of partitioning you want to save the output in.

    - **Autogenerate files (Recommended)**: This is the default value for the number of generated files.
    - **Multiple file output**: Specify the number of file outputs you want. For optimal performance, use the default auto-generated number of files value.
  - **Partition keys**: Choose which columns to use as partitioning keys in the output. To add more partition keys, choose **Add a partition key**.

  File partitioning is not supported for Tableau Hyper as a target format.

Using Data Catalog tables for the data target

For all data sources except Amazon S3 and connectors, a table must exist in the AWS Glue Data Catalog for the target type that you choose. AWS Glue Studio does not create the Data Catalog table.

To configure the data properties for a target that uses a Data Catalog table

1. Go to the visual editor for a new or saved job.
2. Choose a data target node in the job diagram.
3. Choose the **Data target properties** tab, and then enter the following information:
   - **Database**: Choose the database that contains the table you want to use as the target from the list. This database must already exist in the Data Catalog.
   - **Table**: Choose the table that defines the schema of your output data from the list. This table must already exist in the Data Catalog.

     A table in the Data Catalog consists of the names of columns, data type definitions, partition information, and other metadata about the target dataset. Your job writes to a location described by this table in the Data Catalog.

     For more information about creating tables in the Data Catalog, see Defining Tables in the Data Catalog in the *AWS Glue Developer Guide*.
   - **Data catalog update options**
     - **Do not change table definition**: (Default) Choose this option if you don't want the job to update the Data Catalog, even if the schema changes, or new partitions are added.

- **Update schema and add new partitions**: If you choose this option, the job updates the Data Catalog table if the schema changes or new partitions are added.
- **Keep existing schema and add new partitions**: If you choose this option, the job updates the Data Catalog table only to add new partitions.
- **Partition keys**: Choose which columns to use as partitioning keys in the output. To add more partition keys, choose **Add a partition key**.

Using a connector for the data target

If you select a connector for the **Node type**, follow the instructions at Authoring jobs with custom connectors to finish configuring the data target properties.

Editing or uploading a job script

Use the AWS Glue Studio visual editor to edit the job script or upload your own script.

You can use the visual editor to edit job nodes only if the jobs were created with AWS Glue Studio. If the job was created using the AWS Glue console, through API commands, or with the command line interface (CLI), you can use the script editor in AWS Glue Studio to edit the job script, parameters, and schedule. You can also edit the script for a job created in AWS Glue Studio by converting the job to script-only mode.

To edit the job script or upload your own script

1. If creating a new job, on the **Jobs** page, choose the **Spark script editor** option to create a Spark job or choose the **Python Shell script editor** to create a Python shell job. You can either write a new script, or upload an existing script. If you choose **Spark script editor**, you can write or upload either a Scala or Python script. If you choose **Python Shell script editor**, you can only write or upload a Python script.

   After choosing the option to create a new job, in the **Options** section that appears, you can choose to either start with a starter script (**Create a new script with boilerplate code**), or you can upload a local file to use as the job script.

   If you chose **Spark script editor**, you can upload either Python or Scala script files. Scala scripts must have the file extension `.scala`. Python scripts must be recognized as files of type Python. If you chose **Python Shell script editor**, you can upload only Python script files.

When you are finished making your choices, choose **Create** to create the job and open the visual editor.

2. Go to the visual job editor for the new or saved job, and then choose the **Script** tab.
3. If you didn't create a new job using one of the script editor options, and you have never edited the script for an existing job, the **Script** tab displays the heading **Script (Locked)**. This means the script editor is in read-only mode. Choose **Edit script** to unlock the script for editing.

   To make the script editable, AWS Glue Studio converts your job from a visual job to a script-only job. If you unlock the script for editing, you can't use the visual editor anymore for this job after you save it.

   In the confirmation window, choose **Confirm** to continue or **Cancel** to keep the job available for visual editing.

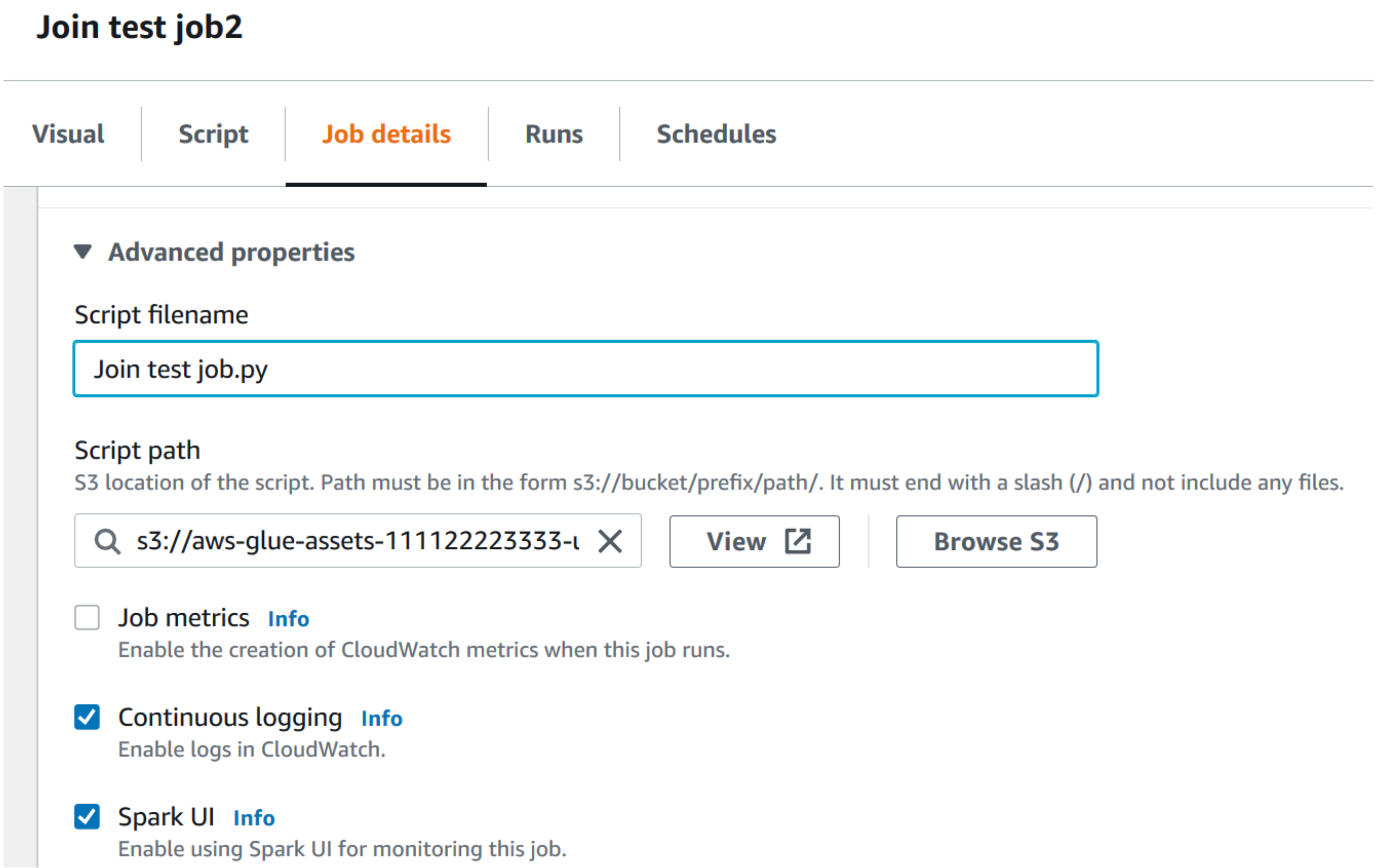
   If you choose **Confirm**, the **Visual** tab no longer appears in the editor. You can use AWS Glue Studio to modify the script using the script editor, modify the job details or schedule, or view job runs.

   > **Note**
   >
   > Until you save the job, the conversion to a script-only job is not permanent. If you refresh the console web page, or close the job before saving it and reopen it in the visual editor, you will still be able to edit the individual nodes in the visual editor.

4. Edit the script as needed.

   When you are done editing the script, choose **Save** to save the job and permanently convert the job from visual to script-only.

5. (Optional) You can download the script from the AWS Glue Studio console by choosing the **Download** button on the **Script** tab. When you choose this button, a new browser window opens, displaying the script from its location in Amazon S3. The **Script filename** and **Script path** parameters in the **Job details** tab of the job determine the name and location of the script file in Amazon S3.



Script filename

Join test job.py

Script path

S3 location of the script. Path must be in the form s3://bucket/prefix/path/. It must end with a slash (/) and not include any files.

s3://aws-glue-assets-111122223333-ı

☐ Job metrics Info

Enable the creation of CloudWatch metrics when this job runs.

☑ Continuous logging Info

Enable logs in CloudWatch.

☑ Spark UI Info

Enable using Spark UI for monitoring this job.

When you save the job, AWS Glue save the job script at the location specified by these fields. If you modify the script file at this location within Amazon S3, AWS Glue Studio will load the modified script the next time you edit the job.

Creating and editing Scala scripts in AWS Glue Studio

When you choose the script editor for creating a job, by default, the job programming language is set to `Python 3`. If you choose to write a new script instead of uploading a script, AWS Glue Studio starts a new script with boilerplate text written in Python. If you want to write a Scala script instead, you must first configure the script editor to use Scala.

> **Note**
>
> If you choose Scala as the programming language for the job and use the visual editor to design your job, the generated job script is written in Scala, and no further actions are needed.

To write a new Scala script in AWS Glue Studio

1. Create a new job by choosing the **Spark script editor** option.
2. Under **Options**, choose **Create a new script with boilerplate code**.
3. Choose the **Job details** tab and set **Language** to `Scala` (instead of `Python 3`).

> **Note**
>
> The **Type** property for the job is automatically set to `Spark` when you choose the **Spark script editor** option to create a job.

4. Choose the **Script** tab.
5. Remove the Python boilerplate text. You can replace it with the following Scala boilerplate text.

```
import com.amazonaws.services.glue.{DynamicRecord, GlueContext}
import org.apache.spark.SparkContext
import com.amazonaws.services.glue.util.JsonOptions
import com.amazonaws.services.glue.util.GlueArgParser
import com.amazonaws.services.glue.util.Job

object MyScript {
  def main(args: Array[String]): Unit = {
    val sc: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(sc)

    }
}
```

6. Write your Scala job script in the editor. Add additional `import` statements as needed.

Creating and editing Python shell jobs in AWS Glue Studio

When you choose the Python shell script editor for creating a job, you can upload an existing Python script, or write a new one. If you choose to write a new script, boilerplate code is added to the new Python job script.

To create a new Python shell job

Refer to the instructions at [Starting jobs in AWS Glue Studio](#).

The job properties that are supported for Python shell jobs are not the same as those supported for Spark jobs. The following list describes the changes to the available job parameters for Python shell jobs on the **Job details** tab.

- The **Type** property for the job is automatically set to `Python Shell` and can't be changed.
- Instead of **Language**, there is a **Python version** property for the job. Currently, Python shell jobs created in AWS Glue Studio use Python 3.6.
- The **Glue version** property is not available, because it does not apply to Python shell jobs.
- Instead of **Worker type** and **Number of workers**, a **Data processing units** property is shown instead. This job property determines how many data processing units (DPUs) are consumed by the Python shell when running the job.
- The **Job bookmark** property is not available, because it is not supported for Python shell jobs.
- Under **Advanced properties**, the following properties are not available for Python shell jobs.
  - **Job metrics**
  - **Continuous logging**
  - **Spark UI** and **Spark UI logs path**
  - **Dependent jars path**, under the heading **Libraries**

Changing the parent nodes for a node in the job diagram

You can change a node's parents to move nodes within the job diagram or to change a data source for a node.

To change the parent node

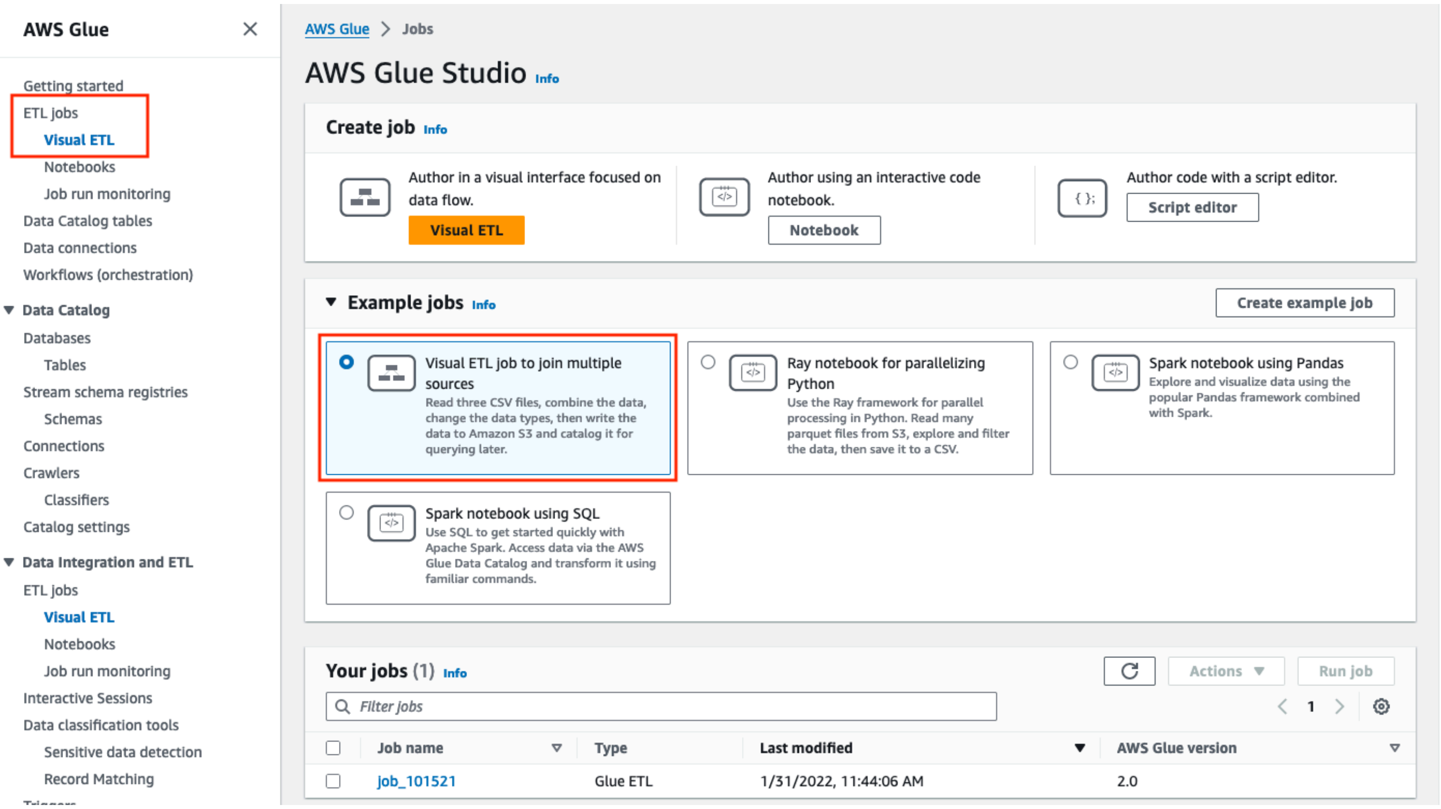
1. Choose the node in the job diagram that you want to modify.
2. In the node details panel, on the **Node properties** tab, under the heading **Node parents** remove the current parent for the node.
3. Choose a new parent node from the list.
4. Modify the other properties of the node as needed to match the newly selected parent node.

If you modified a node by mistake, you can use the **Undo** button on the toolbar to reverse the action.

Deleting nodes from the job diagram

When working with Visual ETL jobs, you can remove nodes from the canvas without having to re-add or restructure any nodes that are connected to the removed node.

In the example below, you can follow along by choosing **ETL jobs > Visual ETL**, then in **Example jobs**, choosing **Visual ETL job to join multiple sources**. Choose **Create example job** to create a job and follow along with the steps below.


AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Interactive Sessions
Data classification tools
Sensitive data detection
Record Matching
AWS Glue > Jobs
AWS Glue Studio Info
Create job Info
Visual ETL
Notebook
Script editor
Example jobs Info
Create example job
Your jobs (1) Info
Actions
Run job
Filter jobs
1

Author in a visual interface focused on data flow.

Author using an interactive code notebook.

Author code with a script editor.

● Visual ETL job to join multiple sources
Read three CSV files, combine the data, change the data types, then write the data to Amazon S3 and catalog it for querying later.

○ Ray notebook for parallelizing Python
Use the Ray framework for parallel processing in Python. Read many parquet files from S3, explore and filter the data, then save it to a CSV.

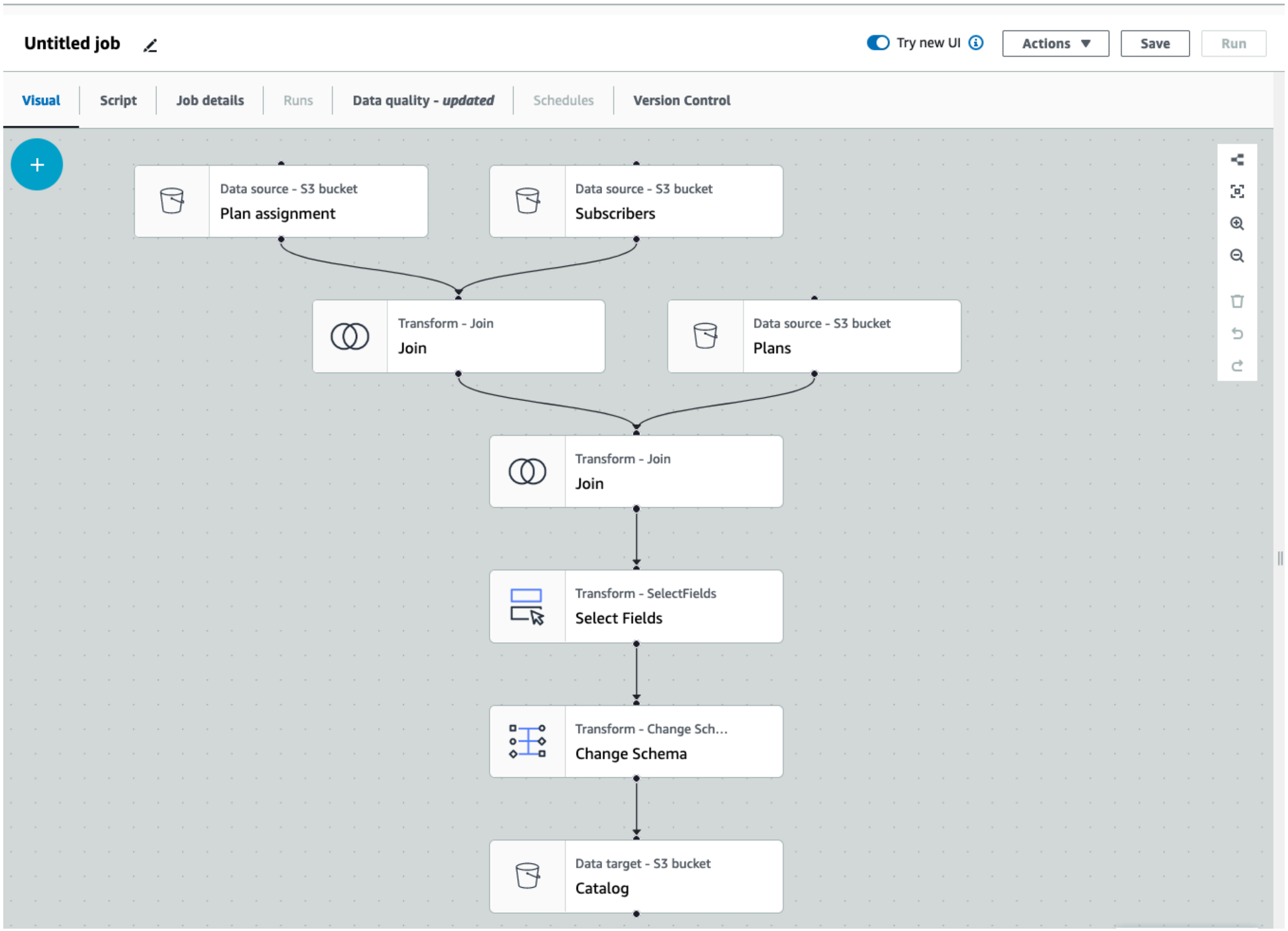
○ Spark notebook using Pandas
Explore and visualize data using the popular Pandas framework combined with Spark.

○ Spark notebook using SQL
Use SQL to get started quickly with Apache Spark. Access data via the AWS Glue Data Catalog and transform it using familiar commands.

☐	Job name	Type	Last modified	AWS Glue version
☐	job_101521	Glue ETL	1/31/2022, 11:44:06 AM	2.0

To remove a node from the canvas

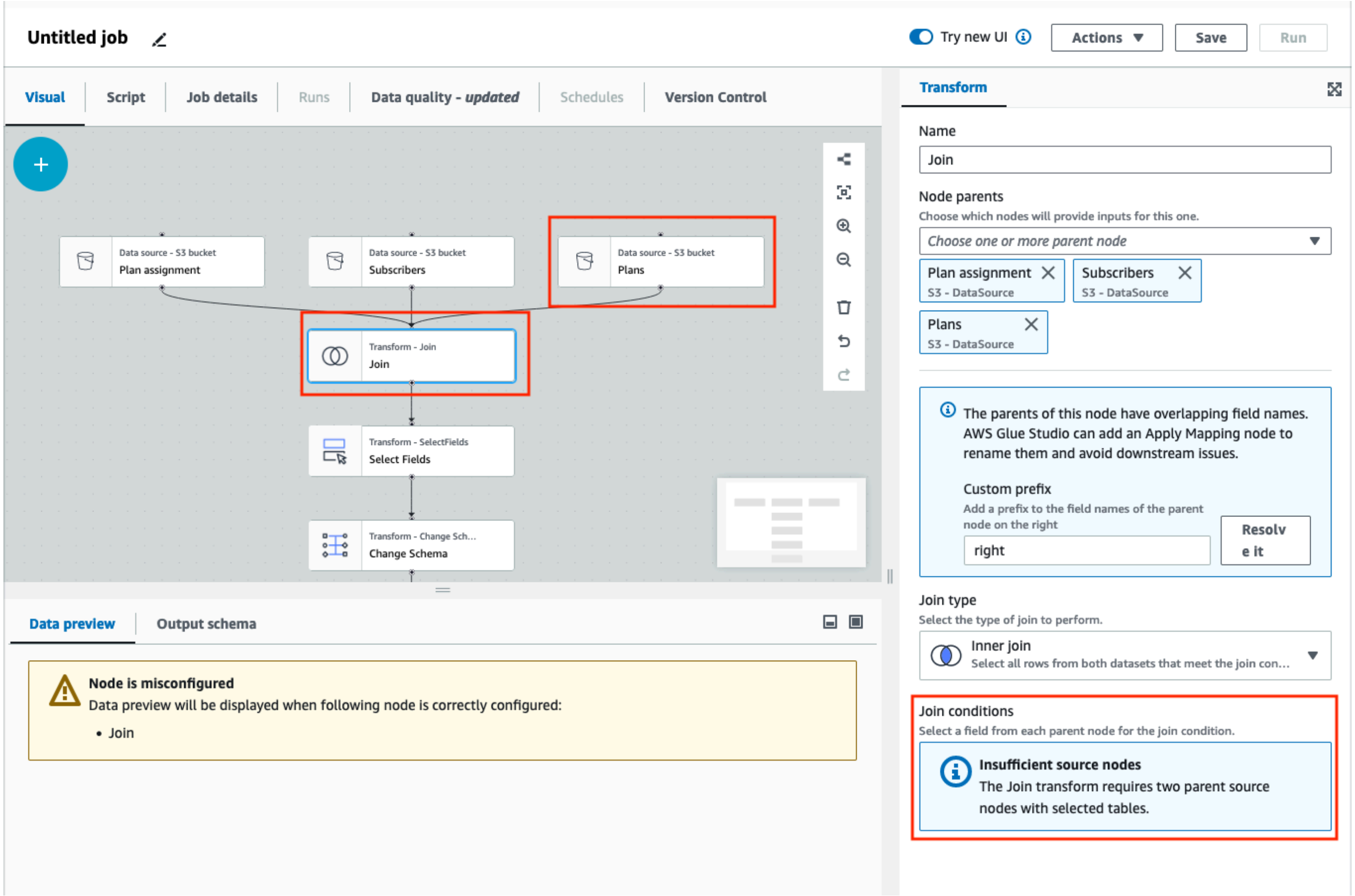
1. From the AWS Glue console, choose **Visual ETL** from the navigation menu and choose an existing job. The job canvas displays the example job as depicted below.


Untitled job
Try new UI
Actions
Save
Run
Visual
Script
Job details
Runs
Data quality - updated
Schedules
Version Control
Data source - S3 bucket
Plan assignment
Data source - S3 bucket
Subscribers
Transform - Join
Join
Data source - S3 bucket
Plans
Transform - Join
Join
Transform - SelectFields
Select Fields
Transform - Change Sch...
Change Schema
Data target - S3 bucket
Catalog

2. Choose the node you want to remove. The canvas will zoom in to the node. In the toolbar on the right side of the canvas, choose the **Trash** icon. This will remove the node and any node connected to the node will move to take its place in the workflow. In this example, the first **Join** node was deleted from the canvas.

   If you delete a node in the workflow, AWS Glue will re-arrange the nodes so that they are organized in a way that does not result in an invalid workflow. You may still need to correct a node's configuration.

   In the example, the **Join** node beneath the **Subscribers** node was removed. As a result, the **Plans** source node has been moved to the top level and is still connected to the child **Join** node. The **Join** node now requires additional configuration since **Join** requires two parent source nodes with selected tables. The **Transform** tab to the right of the canvas displays the missing requirement under **Join conditions** .


Untitled job
Try new UI
Actions
Save
Run
Visual
Script
Job details
Runs
Data quality - updated
Schedules
Version Control
Data source - S3 bucket
Plan assignment
Data source - S3 bucket
Subscribers
Data source - S3 bucket
Plans
Transform - Join
Join
Transform - SelectFields
Select Fields
Transform - Change Sch...
Change Schema
Data preview
Output schema
Transform

Name: Join

Node parents

Choose which nodes will provide inputs for this one.

Choose one or more parent node

Plan assignment S3 - DataSource

Subscribers S3 - DataSource

Plans S3 - DataSource

The parents of this node have overlapping field names. AWS Glue Studio can add an Apply Mapping node to rename them and avoid downstream issues.

Custom prefix

Add a prefix to the field names of the parent node on the right

right

Resolve it

Join type

Select the type of join to perform.

Inner join

Select all rows from both datasets that meet the join con...

Join conditions

Select a field from each parent node for the join condition.

Insufficient source nodes

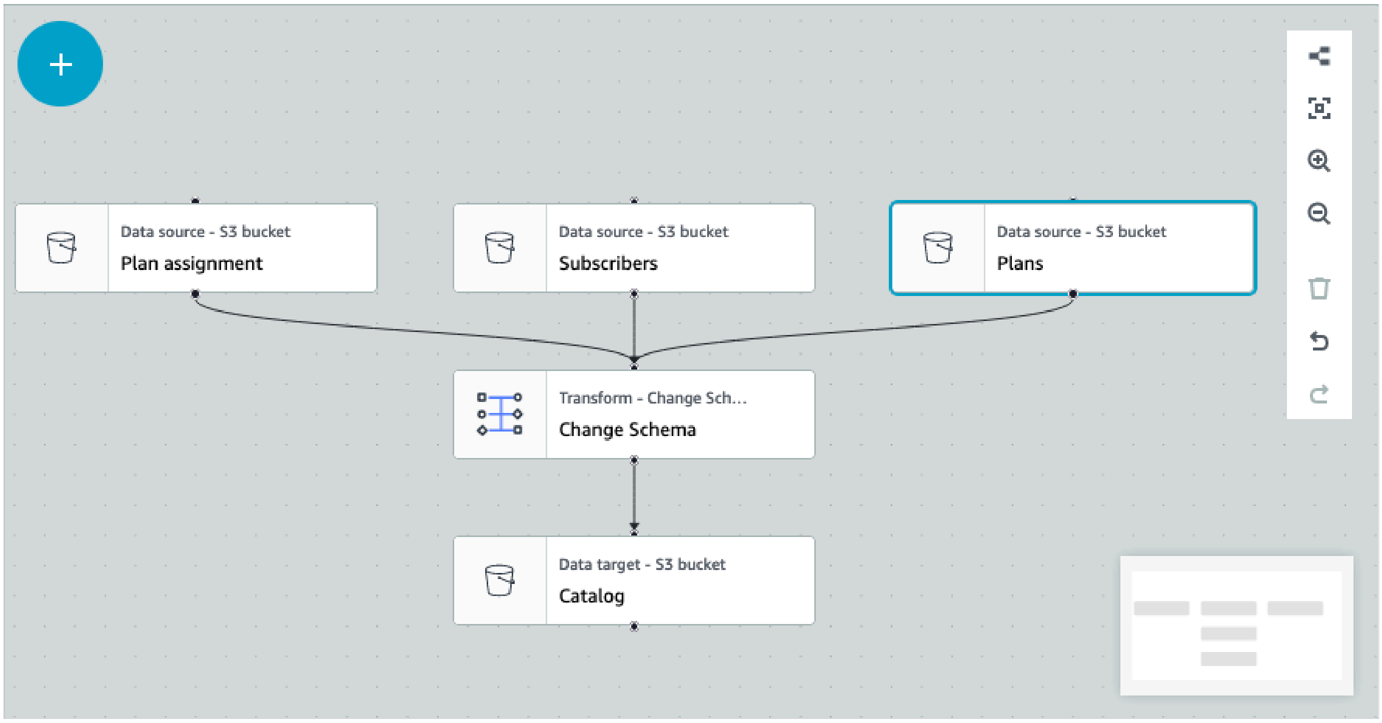
The Join transform requires two parent source nodes with selected tables.

Node is misconfigured

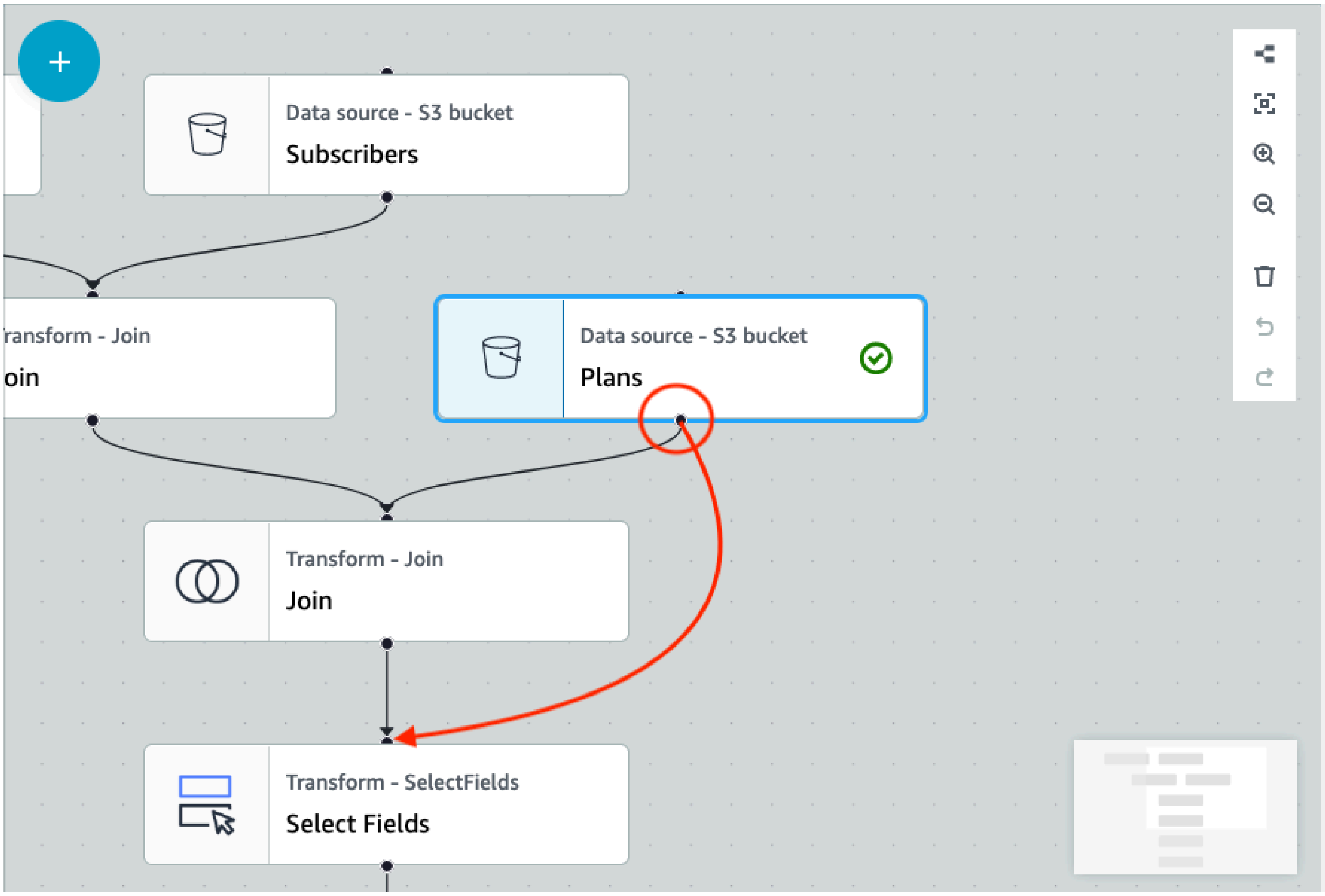
Data preview will be displayed when following node is correctly configured:

- Join

3. Delete the second **Join** node and **Select Fields** node. When the nodes have been deleted, the workflow will look like the example below.


Data source - S3 bucket
Plan assignment
Data source - S3 bucket
Subscribers
Data source - S3 bucket
Plans
Transform - Change Sch...
Change Schema
Data target - S3 bucket
Catalog

4. To modify the node connections, click on the node's handle and drag the connection to a new node. This will allow you to delete nodes and rearrange the nodes in a logical flow. In the example, a new connection is being made by clicking the handle on the Plans node and dragging the connection to the Join node as depicted by the red arrow.


Data source - S3 bucket
Subscribers
ransform - Join
oin
Data source - S3 bucket
Plans
Transform - Join
Join
Transform - SelectFields
Select Fields

5. If you need to undo any action, choose the **Undo** icon directly beneath the **Trash** icon in the toolbar on the right side of the canvas.

Adding source and target parameters to the AWS Glue Data Catalog node

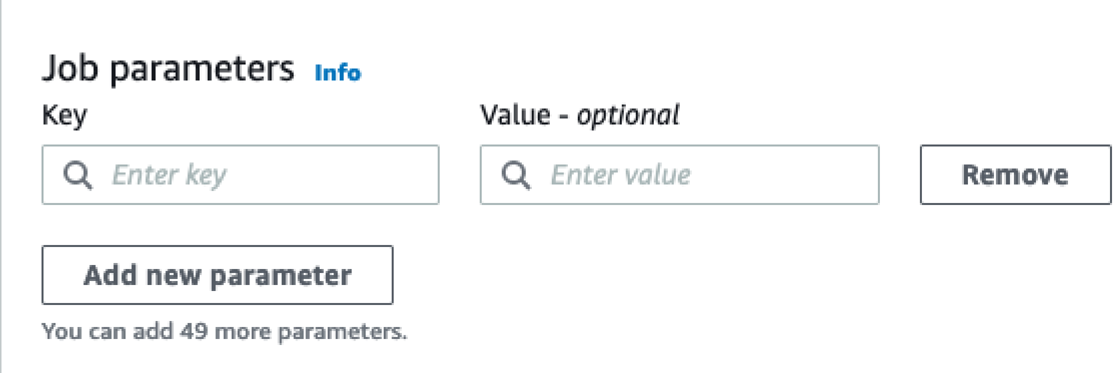
AWS Glue Studio allows you to parameterize visual jobs. Since catalog table names in production and development environment may be different, you can define and select runtime parameters for databases and tables that will run when your job runs.

Job parameterization allows you to parameterize sources and targets, and save those parameters to the job when using the AWS Glue Data Catalog node. When you specify sources and targets as paramters, you are enabling the reusability of jobs, particularly when using the same job in multiple environments. This is useful when promoting code across deployment environments by saving time and effort in managing your sources and targets. In addition, the custom parameters you specify will override any default arguments for specific runs of AWS Glue jobs.

To add source and target parameters

Whether you are using the AWS Glue Data Catalog node as a source or a target, you can define runtime parameters in the **Advanced properties** section on the **Job details** tab.

1. Choose the AWS Glue Data Catalog node as either the source node or the target node.
2. Choose the **Job details** tab.
3. Choose **Advanced properties**.
4. In the Job parameters section, enter a key value. For example, `--db.source` would be the parameter for a database source. You can enter any name for the key, as long as the key name is followed by the 'dash dash'.

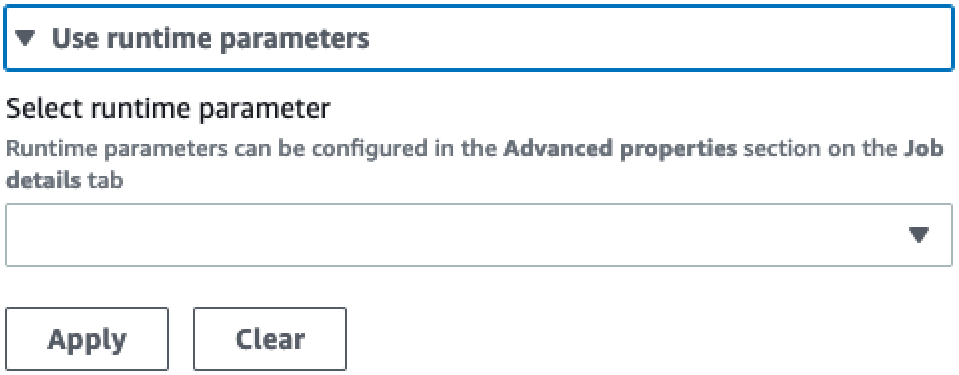


5. Enter the value. For example, `databasename` would be the value for database being parameterized.
6. Choose **Add new parameter** if you want to add more parameters. Max 50 parameters is allowed. Once the key value pair has been defined, you can use the parameter in the AWS Glue Data Catalog node.

To select a runtime parameter

> **Note**
>
> The process to select runtime parameters for databases and tables is the same whether the the AWS Glue Data Catalog node is the source or the target.

1. Choose the AWS Glue Data Catalog node as either the source node or the target node.
2. In the **Data source properties - Data Catalog** tab, under **Database**, choose **Use runtime parameters**.


▼ Use runtime parameters
Select runtime parameter
Runtime parameters can be configured in the **Advanced properties** section on the **Job details** tab
Apply
Clear

3. Choose a parameter from the drop-down menu. For example, when you select a parameter you defined for a source database, the database will automatically populate in the database drop-down menu when you choose **Apply**.
4. In the Table section, choose a parameter you already defined as a source table. When you choose **Apply**, the table is automatically populated as the table to use.
5. When you save and run the job, AWS Glue Studio will reference the selected parameters during the job run.

Using Git version control systems in AWS Glue

Note

Notebooks are not currently supported for version control in AWS Glue Studio. However, version control for AWS Glue job scripts and visual ETL jobs are supported.

If you have remote repositories and want to manage your AWS Glue jobs using your repositories, you can use AWS Glue Studio or the AWS CLI to sync changes to your repositories and your jobs in AWS Glue. When you sync changes this way, you're pushing the job from AWS Glue Studio to your repository, or pulling from the repository to AWS Glue Studio.

With Git integration in AWS Glue Studio, you can:

- Integrate with Git version control systems, such as AWS CodeCommit, GitHub, GitLab, and Bitbucket
- Edit AWS Glue jobs in AWS Glue Studio whether you use visual jobs or script jobs and sync them to a repository

- Parameterize sources and targets in jobs
- Pull jobs from a repository and edit them in AWS Glue Studio
- Test jobs by pulling from branches and/or pushing to branches utilizing multi-branch workflows in AWS Glue Studio
- Download files from a repository and upload jobs into AWS Glue Studio for cross-account job creation
- Use your automation tool of choice (for example, Jenkins, AWS CodeDeploy, etc.)

This video demonstrates how you can integrate AWS Glue with Git and build a continuous and collaborative code pipeline.

IAM permissions

Ensure the job has one of the following IAM permissions. For more information on how to set up IAM permissions, see Set up IAM permissions for AWS Glue Studio.

- `AWSGlueServiceRole`
- `AWSGlueConsoleFullAccess`

At minimum, the following actions are needed for Git integration:

- `glue:UpdateJobFromSourceControl` — to be able to update AWS Glue with a job present in a version control system
- `glue:UpdateSourceControlFromJob` — to be able to update the version control system with a job stored in AWS Glue
- `s3:GetObject` — to be able to retrieve the script for the job while pushing to version control system
- `s3:PutObject` — to be able to update the script when pulling a job from a source control system

Prerequisites

In order to push jobs to a source control repository, you will need:

- a repository that has already been created by your administrator
- a branch in the repository

- a personal access token (for Bitbucket, this is the Repository Access Token)
- the username of the repository owner
- set permissions in the repository to allow AWS Glue Studio to read and write to the repository
  - **GitLab** – set token scopes to api, read_repository, and write_repository
  - **Bitbucket** – set permissions to:
    - **Workspace membership** – read, write
    - **Projects** – write, admin read
    - **Repositories** – read, write, admin, delete

> **Note**
>
> When using AWS CodeCommit, personal access token and repository owner are not needed. See Getting started with Git and AWS CodeCommit.

Using jobs from your source control repository in AWS Glue Studio

In order to pull a job from your source control repository that is not in AWS Glue Studio, and to use that job in AWS Glue Studio, the prerequisites will depend on the type of job.

For visual jobs:

- you need a folder and a JSON file of the job definition that matches the job name

  For example, see the job definition below. The branch in your repository should contain a path `my-visual-job/my-visual-job.json` where both the folder and the JSON file match the job name

```
{
  "name" : "my-visual-job",
  "description" : "",
  "role" : "arn:aws:iam::aws_account_id:role/Rolename",
  "command" : {
    "name" : "glueetl",
    "scriptLocation" : "s3://foldername/scripts/my-visual-job.py",
    "pythonVersion" : "3"
  },
  "codeGenConfigurationNodes" : "{\"node-nodeID\":{\"S3CsvSource\":
{\"AdditionalOptions\":{\"EnableSamplePath\":false,\"SamplePath\":\"s3://notebook-
test-input/netflix_titles.csv\"},\"Escaper\":\"\",\"Exclusions\":[],\"Name\":\"Amazon
 S3\",\"OptimizePerformance\":false,\"OutputSchemas\":[{\"Columns\":[{\"Name\":
\"show_id\",\"Type\":\"string\"},{\"Name\":\"type\",\"Type\":\"string\"},{\"Name\":
\"title\",\"Type\":\"choice\"},{\"Name\":\"director\",\"Type\":\"string\"},{\"Name\":
\"cast\",\"Type\":\"string\"},{\"Name\":\"country\",\"Type\":\"string\"},{\"Name\":
\"date_added\",\"Type\":\"string\"},{\"Name\":\"release_year\",\"Type\":\"bigint\"},
{\"Name\":\"rating\",\"Type\":\"string\"},{\"Name\":\"duration\",\"Type\":\"string
\"},{\"Name\":\"listed_in\",\"Type\":\"string\"},{\"Name\":\"description\",\"Type
\":\"string\"}]}],\"Paths\":[\"s3://dalamgir-notebook-test-input/netflix_titles.csv
\"],\"QuoteChar\":\"quote\",\"Recurse\":true,\"Separator\":\"comma\",\"WithHeader
\":true}}}"
}
```

For script jobs:

- you need a folder, a JSON file of the job definition, and the script
- the folder and JSON file should match the job name. The script name needs to match the `scriptLocation` in the job definition along with the file extension

  For example, in the job definition below, the branch in your repository should contain a path `my-script-job/my-script-job.json` and `my-script-job/my-script-job.py`. The script name should match the name in the `scriptLocation` including the extension of the script

```
{
  "name" : "my-script-job",
  "description" : "",
  "role" : "arn:aws:iam::aws_account_id:role/Rolename",
  "command" : {
    "name" : "glueetl",
    "scriptLocation" : "s3://foldername/scripts/my-script-job.py",
    "pythonVersion" : "3"
  }
}
```

Limitations

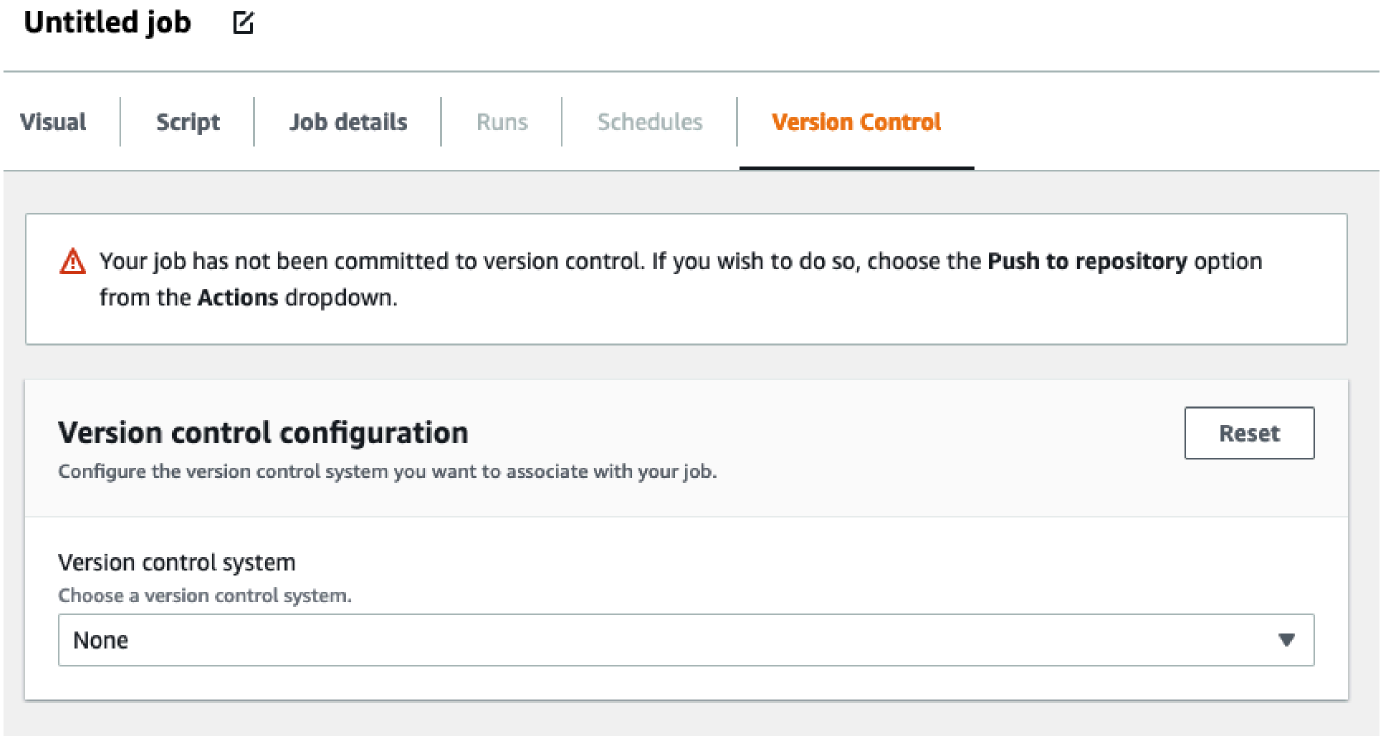
- AWS Glue currently does not support pushing/pulling from GitLab-Groups.

Connecting version control repositories with AWS Glue

You can enter your version control repository details and manage them in the **Version Control** tab in the AWS Glue Studio job editor. To integrate with your Git repository, you must connect to your repository every time you log in to AWS Glue Studio.

To connect a Git version control system:

1. In AWS Glue Studio, start a new job and choose the **Version Control** tab.



Your job has not been committed to version control. If you wish to do so, choose the **Push to repository** option from the **Actions** dropdown.

Version control configuration

Configure the version control system you want to associate with your job.

Version control system

Choose a version control system.

None

2. In **Version control system**, choose the Git Service from the available options by clicking on the drop-down menu.
   - AWS CodeCommit
   - GitHub
   - GitLab
   - Bitbucket

3. Depending on the Git version control system you choose, you will have different fields to complete.

   For AWS CodeCommit:

   Complete the repository configuration by selecting the repository and branch for your job:

   - **Repository** — if you have set up repositories in AWS CodeCommit, select the repository from the drop-down menu. Your repositories will automatically populate in the list
   - **Branch** — select the branch from the drop-down menu
   - **Folder** — *optional* - enter the name of the folder in which to save your job. If left empty, a folder is automatically created. The folder name defaults to the job name

   For GitHub:

   Complete the GitHub configuration by completing the fields:

   - **Personal access token** — this is the token provided by the GitHub repository. For more information on personal access tokens, see GitHub Docs
   - **Repository owner** — this is the owner of the GitHub repository.

   Complete the repository configuration by selecting the repository and branch from GitHub.

   - **Repository** — if you have set up repositories in GitHub, select the repository from the drop-down menu. Your repositories will automatically populate in the list
   - **Branch** — select the branch from the drop-down menu
   - **Folder** — *optional* - enter the name of the folder in which to save your job. If left empty, a folder is automatically created. The folder name defaults to the job name

   For GitLab:

   > **Note**
   >
   > AWS Glue currently does not support pushing/pulling from GitLab-Groups.

- **Personal access token** — this is the token provided by the GitLab repository. For more information on personal access tokens, see GitLab Personal access tokens
- **Repository owner** — this is the owner of the GitLab repository.

Complete the repository configuration by selecting the repository and branch from GitLab.

- **Repository** — if you have set up repositories in GitLab, select the repository from the drop-down menu. Your repositories will automatically populate in the list
- **Branch** — select the branch from the drop-down menu
- **Folder** — *optional* - enter the name of the folder in which to save your job. If left empty, a folder is automatically created. The folder name defaults to the job name

For Bitbucket:

- **App password** — Bitbucket uses App passwords and not Repository Access Tokens. For more information on App passwords, see App passwords .
- **Repository owner** — this is the owner of the Bitbucket repository. In Bitbucket, the owner is the creator of the repository.

Complete the repository configuration by selecting the workspace, repository, branch, and folder from Bitbucket.

- **Workspace** – if you have workspaces set up in Bitbucket, select the workspace from the drop-down menu. Your workspaces are automatically populated
- **Repository** — if you have set up repositories in Bitbucket, select the repository from the drop-down menu. Your repositories are automatically populated
- **Branch** — select the branch from the drop-down menu. Your branches are automatically populated
- **Folder** — *optional* - enter the name of the folder in which to save your job. If left empty, a folder is automatically created with the job name.

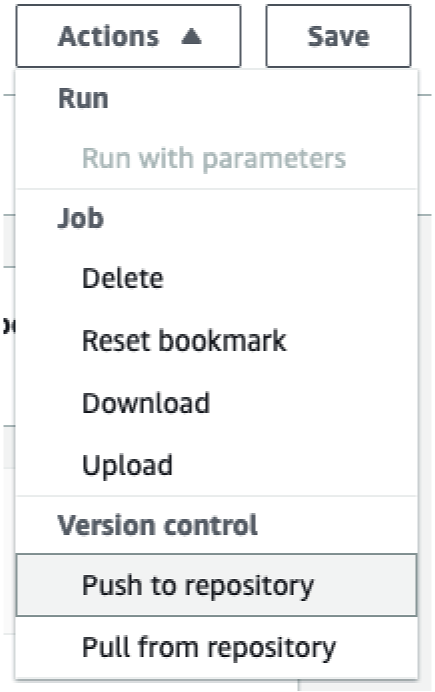
4. Choose **Save** at the top of the AWS Glue Studio job

Pushing AWS Glue jobs to the source repository

Once you've entered the details of your version control system, you can edit jobs in AWS Glue Studio and push the jobs to your source repository. If you're unfamiliar with Git concepts such as pushing and pulling, see this tutorial on Getting started with Git and AWS CodeCommit.

In order to push your job to a repository, you need to enter the details of your version control system and save your job.

1. In the AWS Glue Studiojob, choose **Actions**. This will open additional menu options.


Actions ▲
Save
Run
Run with parameters
Job
Delete
Reset bookmark
Download
Upload
Version control
Push to repository
Pull from repository

2. Choose **Push to repository**.

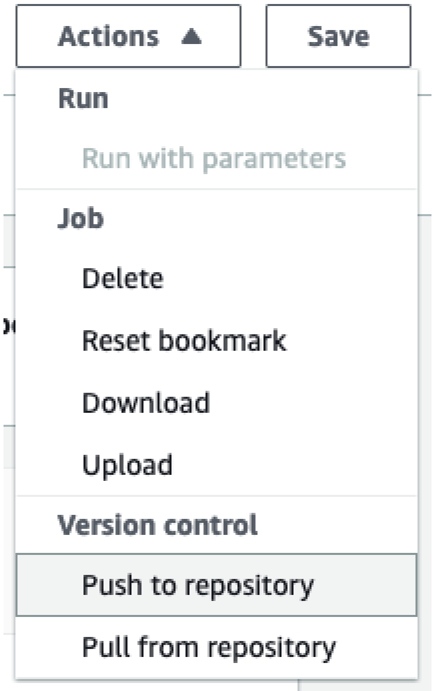
   This action will save the job. When you push to repository, AWS Glue Studio pushes the last saved change. If the job in the repository was modified by you or another user and is out of sync with the job in AWS Glue Studio, the job in the repository is overwritten with the job saved in AWS Glue Studio when you push the job from AWS Glue Studio.

3. Choose **Confirm** to complete the action. This creates a new commit in the repository. If you are using AWS CodeCommit, a confirmation message will display a link to the latest commit on AWS CodeCommit.

Pulling AWS Glue jobs from the source repository

Once you've entered details of your Git repository into the **Version control** tab, you can also pull jobs from your repository and edit them in AWS Glue Studio.

1. In the AWS Glue Studio job, choose **Actions**. This will open additional menu options.


Actions
Save
Run
Run with parameters
Job
Delete
Reset bookmark
Download
Upload
Version control
Push to repository
Pull from repository

2. Choose **Pull from repository**.
3. Choose **Confirm**. This takes the latest commit from the repository and updates your job in AWS Glue Studio.
4. Edit your job in AWS Glue Studio. If you make changes, you can sync your job to your repository by choosing **Push to repository** from the **Actions** drop-down menu.

Authoring code with AWS Glue Studio notebooks

Data engineers can author AWS Glue jobs faster and more easily than before using the interactive notebook interface in AWS Glue Studio or interactive sessions in AWS Glue.

Limitations

- AWS Glue Studio notebooks do not support Scala.

Topics

- Overview of using notebooks
- Creating an ETL job using notebooks in AWS Glue Studio
- Notebook editor components
- Saving your notebook and job script
- Managing notebook sessions
- Using Amazon Q Developer with AWS Glue Studio notebooks

Overview of using notebooks

AWS Glue Studio allows you to interactively author jobs in a notebook interface based on Jupyter Notebooks. Through notebooks in AWS Glue Studio, you can edit job scripts and view the output without having to run a full job, and you can edit data integration code and view the output without having to run a full job, and you can add markdown and save notebooks as .ipynb files and job scripts. You can start a notebook without installing software locally or managing servers. When you are satisfied with your code, AWS Glue Studio can convert your notebook to a Glue job with the click of a button.

Some benefits of using notebooks include:

- No cluster to provision or manage
- No idle clusters to pay for
- No up-front configuration required
- No installation of Jupyter notebooks required
- The same runtime/platform as AWS Glue ETL

When you start a notebook through AWS Glue Studio, all the configuration steps are done for you so that you can explore your data and start developing your job script after only a few seconds. AWS Glue Studio configures a Jupyter notebook with the AWS Glue Jupyter kernel. You don’t have to configure VPCs, network connections, or development endpoints to use this notebook.

To create jobs using the notebook interface:

- configure the necessary IAM permissions.
- start a notebook session to create a job

- write code in the cells in the notebook
- run and test the code to view the output
- save the job

After your notebook is saved, your notebook is a full AWS Glue job. You can manage all aspects of the job, such as scheduling jobs runs, setting job parameters, and viewing the job run history right along side your notebook.

Creating an ETL job using notebooks in AWS Glue Studio

To start using notebooks in the AWS Glue Studio console

1. Attach AWS Identity and Access Management policies to the AWS Glue Studio user and create an IAM role for your ETL job and notebook.
2. Configure additional IAM security for notebooks, as described in Granting permissions for the IAM role.
3. Open the AWS Glue Studio console at https://console.aws.amazon.com/gluestudio/.

   > **Note**
   >
   > Check that your browser does not block third-party cookies. Any browser that blocks third party cookies either by default or as a user-enabled setting will prevent notebooks from launching. For more information on managing cookies, see:

   - Chrome
   - Firefox
   - Safari
4. Choose the **Jobs** link in the left-side navigation menu.
5. Choose **Jupyter notebook** and then choose **Create** to start a new notebook session.
6. On the **Create job in Jupyter notebook** page, provide the job name, and choose the IAM role to use. Choose **Create job**.

   After a short time period, the notebook editor appears.
7. After you add the code you must execute the cell to initiate a session. There are multiple ways to execute the cell:

    - Press the play button.
    - Use a keyboard shortcut:
        - On MacOS, **Command** + **Enter** to run the cell.
        - On Windows, **Shift** + **Enter** to run the cell.

   For information about writing code using a Jupyter notebook interface, see *The Jupyter Notebook User Documentation* .

8. To test your script, run the entire script, or individual cells. Any command output will be displayed in the area beneath the cell.
9. After you have finished developing your notebook, you can save the job and then run it. You can find the script in the **Script** tab. Any magics you added to the notebook will be stripped away and won't be saved as part of the script of the generated AWS Glue job. AWS Glue Studio will auto-add a `job.commit()` to the end of your generated script from the notebook contents.

   For more information about running jobs, see Start a job run.

Notebook editor components

The notebook editor interface has the following main sections.

- Notebook interface (main panel) and toolbar
- Job editing tabs

The notebook editor

The AWS Glue Studio notebook editor is based on the Jupyter Notebook Application. The AWS Glue Studio notebook interface is similar to that provided by Juypter Notebooks, which is described in the section Notebook user interface . The notebook used by interactive sessions is a Jupyter Notebook.

Although the AWS Glue Studio notebook is similar to Juptyer Notebooks, it differs in a few key ways:

- currently, the AWS Glue Studio notebook cannot install extensions

- you cannot use multiple tabs; there is a 1:1 relationship between a job and a notebook
- the AWS Glue Studio notebook does not have the same top file menu that exists in Jupyter Notebooks
- currently, the AWS Glue Studio notebook only runs with the AWS Glue kernel. Note that you cannot update the kernel on your own.

AWS Glue Studio job editing tabs

The tabs that you use to interact with the ETL job are at the top of the notebook page. They are similar to tabs that appear in the visual job editor of AWS Glue Studio, and they perform the same actions.

- **Notebook** – Use this tab to view the job script using the notebook interface.
- **Job details** – Configure the environment and properties for the job runs.
- **Runs** – View information about previous runs of this job.
- **Schedules** – Configure a schedule for running your job at specific times.

Saving your notebook and job script

You can save your notebook and the job script you are creating at any time. Simply choose the **Save** button in the upper right corner, the same as if you were using the visual or script editor.

When you choose **Save**, the notebook file is saved in the default locations:

- By default, the job script is saved to the Amazon S3 location indicated in the **Job Details** tab, under **Advanced properties**, in the Job details property **Script path**. Job scripts are saved in a subfolder named `Scripts`.
- By default, the notebook file (`.ipynb`) is saved to the Amazon S3 location indicated in the **Job Details** tab, under **Advanced properties**, in the Job details **Script path**. Notebook files are saved in a subfolder named `Notebooks`.

Note

When you save the job, the job script contains only the code cells from the notebook. The Markdown cells and magics aren't included in the job script. However, the `.ipynb` file will contain any markdown and magics.

After you save the job, you can then run the job using the script that you created in the notebook.

Managing notebook sessions

Notebooks in AWS Glue Studio are based on the interactive sessions feature of AWS Glue. There is a cost for using interactive sessions. To help manage your costs, you can monitor the sessions created for your account, and configure the default settings for all sessions.

Change the default timeout for all notebook sessions

By default, the provisioned AWS Glue Studio notebook times out after 12 hours if the notebook was launched and no cells have been executed. There is no cost associated to it and the timeout is not configurable.

Once you execute a cell this will start an interactive session. This session has a default timeout of 48 hours. This timeout can be configured by passing an `%idle_timeout` magic before executing a cell.

To modify the default session timeout for notebooks in AWS Glue Studio

1. In the notebook, enter the `%idle_timeout` magic in a cell and specify the timeout value in minutes.
2. For example: `%idle_timeout 15` will change the default timeout to 15 minutes. If the session is not used in 15 minutes, the session is automatically stopped.

Installing additional Python modules

If you would like to install additional modules to your session using pip you can do so by using `%additional_python_modules` to add them to your session:

```
%additional_python_modules awswrangler, s3://amzn-s3-demo-bucket/mymodule.whl
```

All arguments to additional_python_modules are passed to `pip3 install -m <>`

For a list of available Python modules, see Using Python libraries with AWS Glue.

Changing AWS Glue Configuration

You can use magics to control AWS Glue job configuration values. If you want to change a job configuration value you have to use the proper magic in the notebook. See Magics supported by AWS Glue interactive sessions for Jupyter.

> **Note**
>
> Overriding properties for a running session is no longer available. In order to change the session's configurations, you can stop the session, set the new configurations and then start a new session.

AWS Glue supports various worker types. You can set the worker type with `%worker_type`. For example: `%worker_type G.2X` . Available worker types include G.1X, G.2X, G.4X, G.8X, G.12X, G.16X, R.1X, R.2X, R.4X, and R.8X. The default is G.1X.

You can also specify the Number of workers with `%number_of_workers`. For example, to specify 40 workers: `%number_of_workers 40`.

For more information see Defining Job Properties

Stop a notebook session

To stop a notebook session, use the magic `%stop_session`.

If you navigate away from the notebook in the AWS console, you will receive a warning message where you can choose to stop the session.

Using Amazon Q Developer with AWS Glue Studio notebooks

AWS Glue Studio allows you to interactively author jobs in a notebook interface based on Jupyter Notebooks. Using Amazon Q Developer improves the authoring experience within AWS Glue Studio notebooks.

The Amazon Q Developer extension supports writing code by generating code recommendations and suggesting improvements related to code issues. Amazon Q Developer supports both Python and Scala, the two languages used for coding ETL scripts for Spark jobs in AWS Glue Studio notebooks.

What is Amazon Q Developer?

Amazon Q Developer is a service powered by machine learning that helps improve developer productivity. Amazon Q Developer achieves this by generating code recommendations based on developers' comments in natural language and their code in the IDE. The service integrates with JupyterLab, Amazon SageMaker AI Studio, Amazon SageMaker AI notebook instances, and other integrated development environments (IDEs).

For more information, see Using Amazon Q Developer with AWS Glue Studio.

AWS Glue job run statuses on the console

You can view the status of an AWS Glue extract, transform, and load (ETL) job while it is running or after it has stopped. You can view the status using the AWS Glue console.

Accessing the job monitoring dashboard

You access the job monitoring dashboard by choosing the **Job run monitoring** link in the AWS Glue navigation pane under **ETL jobs**.

Overview of the job monitoring dashboard

The job monitoring dashboard provides an overall summary of the job runs, with totals for the jobs with a status of **Running**, **Canceled**, **Success**, or **Failed**. Additional tiles provide the overall job run success rate, the estimated DPU usage for jobs, a breakdown of the job status counts by job type, worker type, and by day.

The graphs in the tiles are interactive. You can choose any block in a graph to run a filter that displays only those jobs in the **Job runs** table at the bottom of the page.

You can change the date range for the information displayed on this page by using the **Date range** selector. When you change the date range, the information tiles adjust to show the values for the specified number of days before the current date. You can also use a specific date range if you choose **Custom** from the date range selector.

Job runs view

Note

Job run history is accessible for 90 days for your workflow and job run.

The **Job runs** resource list shows the jobs for the specified date range and filters.

You can filter the jobs on additional criteria, such as status, worker type, job type, and the job name. In the filter box at the top of the table, you can enter the text to use as a filter. The table results are updated with rows that contain matching text as you enter the text.

You can view a subset of the jobs by choosing elements from the graphs on the job monitoring dashboard. For example, if you choose the number of running jobs in the **Job runs summary** tile, then the **Job runs** list displays only the jobs that currently have a status of `Running`. If you choose one of the bars in the **Worker type breakdown** bar chart, then only job runs with the matching worker type and status are shown in the **Job runs** list.

The **Job runs** resource list displays the details for the job runs. You can sort the rows in the table by choosing a column heading. The table contains the following information:

Property	Description
Job name	The name of the job.
Type	The type of job environment: • **Glue ETL**: Runs in an Apache Spark environment managed by AWS Glue. • **Glue Streaming**: Runs in an Apache Spark environment and performs ETL on data streams. • **Python shell**: Runs Python scripts as a shell.
Start time	The date and time at which this job run was started.
End time	The date and time that this job run completed.

Property	Description
Run status	The current state of the job run. Values can be: • `STARTING` • `RUNNING` • `STOPPING` • `STOPPED` • `SUCCEEDED` • `FAILED` • `TIMEOUT`
Run time	The amount of time that the job run consumed resources.
Capacity	The number of AWS Glue data processing units (DPUs) that were allocated for this job run. For more information about capacity planning, see Monitoring for DPU Capacity Planning in the *AWS Glue Developer Guide*.

Property	Description
Worker type	The type of predefined worker that was allocated when the job ran. Values can be `G.1X`, `G.2X`, `G.4X` or `G.8X`. - **`G.1X` –** When you choose this type, you also provide a value for **Number of workers**. Each worker maps to 1 DPU (4 vCPUs, 16 GB of memory) with 84GB disk (approximately 34GB free). We recommend this worker type for memory-intensive jobs. This is the default **Worker type** for AWS Glue Version 2.0 or later jobs. - **`G.2X` –** When you choose this type, you also provide a value for **Number of workers**. Each worker maps to 2 DPU (8 vCPUs, 32 GB of memory) with 128GB disk (approximately 77GB free). We recommend this worker type for memory-intensive jobs and jobs that run machine learning transforms. - **`G.4X` –** When you choose this type, you also provide a value for **Number of workers**. Each worker maps to 4 DPU (16 vCPUs, 64 GB of memory) with 256GB disk (approximately 235GB free). We recommend this worker type for jobs whose workloads contain your most demanding transforms, aggregations, joins, and queries. This worker type is available only for AWS Glue version 3.0 or later Spark ETL jobs in the following AWS Regions: US East (Ohio), US East (N. Virginia), US West (Oregon), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Canada (Central), Europe

Property	Description
	(Frankfurt), Europe (Ireland), and Europe (Stockholm). - **`G.8X` –** When you choose this type, you also provide a value for **Number of workers**. Each worker maps to 8 DPU (32 vCPUs, 128 GB of memory) with 512GB disk (approximately 487GB free). We recommend this worker type for jobs whose workloads contain your most demanding transforms, aggregations, joins, and queries. This worker type is available only for AWS Glue version 3.0 or later Spark ETL jobs, in the same AWS Regions as supported for the `G.4X` worker type.
DPU hours	The estimated number of DPUs used for the job run. A DPU is a relative measure of processing power. DPUs are used to determine the cost of running your job. For more information, see the AWS Glue pricing page.

You can choose any job run in the list and view additional information. Choose a job run, and then do one of the following:

- Choose the **Actions** menu and the **View job** option to view the job in the visual editor.
- Choose the **Actions** menu and the **Stop run** option to stop the current run of the job.
- Choose the **View CloudWatch logs** button to view the job run logs for that job.
- Choose **View details** to view the job run details page.

Viewing the job run logs

You can view the job logs in a variety of ways:

- On the **Monitoring** page, in the **Job runs** table, choose a job run, and then choose **View CloudWatch logs**.
- In the visual job editor, on the **Runs** tab for a job, choose the hyperlinks to view the logs:
  - **Logs** – Links to the Apache Spark job logs written when continuous logging is enabled for a job run. When you choose this link, it takes you to the Amazon CloudWatch logs in the `/aws-glue/jobs/logs-v2` log group. By default, the logs exclude non-useful Apache Hadoop YARN heartbeat and Apache Spark driver or executor log messages. For more information about continuous logging, see Continuous Logging for AWS Glue Jobs in the *AWS Glue Developer Guide*.
  - **Error logs** – Links to the logs written to `stderr` for this job run. When you choose this link, it takes you to the Amazon CloudWatch logs in the `/aws-glue/jobs/error` log group. You can use these logs to view details about any errors that were encountered during the job run.
  - **Output logs** – Links to the logs written to `stdout` for this job run. When you choose this link, it takes you to the Amazon CloudWatch logs in the `/aws-glue/jobs/output` log group. You can use these logs to see all the details about the tables that were created in the AWS Glue Data Catalog and any errors that were encountered.

Viewing the details of a job run

You can choose a job in the **Job runs** list on the **Monitoring** page, and then choose **View run details** to see detailed information for that run of the job.

The information displayed on the job run detail page includes:

Property	Description
Job name	The name of the job.
Run Status	The current state of the job run. Values can be: • `STARTING` • `RUNNING` • `STOPPING` • `STOPPED` • `SUCCEEDED` • `FAILED`

Property	Description
	- `TIMEOUT`
Glue version	The AWS Glue version used by the job run.
Recent attempt	The number of automatic retry attempts for this job run.
Start time	The date and time at which this job run was started.
End time	The date and time that this job run completed.
Start-up time	The amount of time spent preparing to run the job.
Execution time	The amount of time spent running the job script.
Trigger name	The name of the trigger associated with the job.
Last modified on	The date when the job was last modified.
Security configuration	The security configuration for the job, which includes Amazon S3 encryption, CloudWatch encryption, and job bookmarks encryption settings.
Timeout	The job run timeout threshold value.
Allocated capacity	The number of AWS Glue data processing units (DPUs) that were allocated for this job run. For more information about capacity planning, see Monitoring for DPU Capacity Planning in the *AWS Glue Developer Guide*.
Max capacity	The maximum capacity available to the job run.

Property	Description
Number of workers	The number of workers used for the job run.
Worker type	The type of predefined workers allocated for the job run. Values can be `G.1X` or `G.2X`. • **`G.1X`** – When you choose this type, you also provide a value for **Number of workers**. Each worker maps to 1 DPU (4 vCPUs, 16 GB of memory, 64 GB disk), and provides 1 executor per worker. We recommend this worker type for memory-intensive jobs. This is the default **Worker type** for AWS Glue Version 2.0 or later jobs. • **`G.2X`** – When you choose this type, you also provide a value for **Number of workers**. Each worker maps to 2 DPUs (8 vCPUs, 32 GB of memory, 128 GB disk), and provides 1 executor per worker. We recommend this worker type for memory-intensive jobs and jobs that run machine learning transforms.
Logs	A link to the job logs for continuous logging (`/aws-glue/jobs/logs-v2`).
Output Logs	A link to the job output log files (`/aws-glue/jobs/output`).
Error logs	A link to the job error log files (`/aws-glue/jobs/error`).

You can also view the following additional items, which are available when you view information for recent job runs. For more information, see the section called "View information for recent job runs".

- **Input arguments**

- **Continuous logs**
- **Metrics** – You can see visualizations of basic metrics. For more information on included metrics, see the section called "Viewing Amazon CloudWatch metrics for a Spark job run".
- **Spark UI** – You can visualize Spark logs for your job in the Spark UI. For more information about using the Spark Web UI, see the section called "Monitoring with the Spark UI". Enable this feature by following the procedure in the section called "Enabling the Spark UI for jobs".

Viewing Amazon CloudWatch metrics for a Spark job run

On the details page for a job run, below the **Run details** section, you can view the job metrics. AWS Glue Studio sends job metrics to Amazon CloudWatch for every job run.

AWS Glue reports metrics to Amazon CloudWatch every 30 seconds. The AWS Glue metrics represent delta values from the previously reported values. Where appropriate, metrics dashboards aggregate (sum) the 30-second values to obtain a value for the entire last minute. However, the Apache Spark metrics that AWS Glue passes on to Amazon CloudWatch are generally absolute values that represent the current state at the time they are reported.

> **Note**
>
> You must configure your account to access Amazon CloudWatch, .

The metrics provide information about your job run, such as:

- **ETL Data Movement** – The number of bytes read from or written to Amazon S3.
- **Memory Profile: Heap used** – The number of memory bytes used by the Java virtual machine (JVM) heap.
- **Memory Profile: heap usage** – The fraction of memory (scale: 0–1), shown as a percentage, used by the JVM heap.
- **CPU Load** – The fraction of CPU system load used (scale: 0–1), shown as a percentage.

Viewing Amazon CloudWatch metrics for a Ray job run

On the details page for a job run, below the **Run details** section, you can view the job metrics. AWS Glue Studio sends job metrics to Amazon CloudWatch for every job run.

AWS Glue reports metrics to Amazon CloudWatch every 30 seconds. The AWS Glue metrics represent delta values from the previously reported values. Where appropriate, metrics dashboards aggregate (sum) the 30-second values to obtain a value for the entire last minute. However, the Apache Spark metrics that AWS Glue passes on to Amazon CloudWatch are generally absolute values that represent the current state at the time they are reported.

> **Note**
>
> You must configure your account to access Amazon CloudWatch, as described in .

In Ray jobs, you can view the following aggregated metric graphs. With these, you can build a profile of your cluster and tasks, and can access detailed information about each node. The time-series data that back these graphs is available in CloudWatch for further analysis.

Task Profile: Task State

Shows the number of Ray tasks in the system. Each task lifecycle is given its own time series.

Task Profile: Task Name

Shows the number of Ray tasks in the system. Only pending and active tasks are shown. Each type of task (by name) is given its own time series.

Cluster Profile: CPUs in use

Shows the number of CPU cores that are used. Each node is given its own time series. Nodes are identified by IP addresses, which are ephemeral and only used for identification.

Cluster Profile: Object store memory use

Shows memory use by the Ray object cache. Each memory location (physical memory, cached on disk, and spilled in Amazon S3) is given its own time series. The object store manages data storage across all nodes in the cluster. For more information, see Objects in the Ray documentation.

Cluster Profile: Node count

Shows the number of nodes provisioned for the cluster.

Node Detail: CPU use

Shows CPU utilization on each node as a percentage. Each series shows an aggregated percentage of CPU usage across all cores on the node.

Node Detail: Memory use

Shows memory use on each node in GB. Each series shows memory aggregated between all processes on the node, including Ray tasks and the Plasma store process. This will not reflect objects stored to disk or spilled to Amazon S3.

Node Detail: Disk use

Shows disk use on each node in GB.

Node Detail: Disk I/O speed

Shows disk I/O on each node in KB/s.

Node Detail: Network I/O throughput

Shows network I/O on each node in KB/s.

Node Detail: CPU use by Ray component

Shows CPU use in fractions of a core. Each ray component on each node is given its own time series.

Node Detail: Memory use by Ray component

Shows memory use in GiB. Each ray component on each node is given its own time series.

Detect and process sensitive data

The Detect PII transform identifies Personal Identifiable Information (PII) in your data source. You choose the PII entity to identify, how you want the data to be scanned, and what to do with the PII entity that have been identified by the Detect PII transform.

The Detect PII transform provides the ability to detect, mask, or remove entities that you define, or are pre-defined by AWS. This enables you to increase compliance and reduce liability. For example, you may want to ensure that no personally identifiable information exists in your data that can be read and want to mask social security numbers with a fixed string (such as xxx-xx-xxxx), phone numbers, or addresses.

To work with sensitive data outside of AWS Glue Studio, see Using Sensitive Data Detection outside AWS Glue Studio

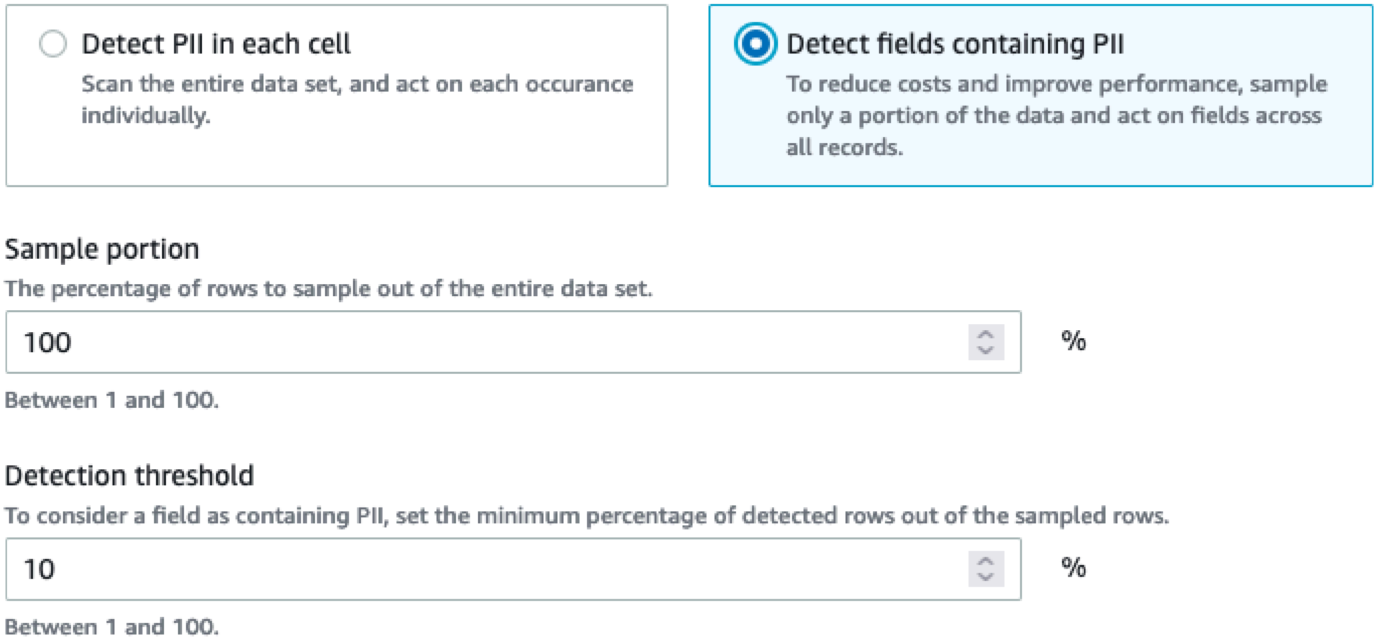
Topics

- Choosing how you want the data to be scanned

- Choosing the PII entities to detect
- Specifying the level of detection sensitivity
- Choosing what to do with identified PII data
- Adding fine-grained action overrides

Choosing how you want the data to be scanned

When you scan your dataset for sensitive data like personally identifiable information (PII), you can choose to detect PII in each row or detect the columns that contain PII data.



○ Detect PII in each cell
Scan the entire data set, and act on each occurance individually.

● Detect fields containing PII
To reduce costs and improve performance, sample only a portion of the data and act on fields across all records.

Sample portion
The percentage of rows to sample out of the entire data set.
100 %
Between 1 and 100.

Detection threshold
To consider a field as containing PII, set the minimum percentage of detected rows out of the sampled rows.
10 %
Between 1 and 100.

When you choose **Detect PII in each cell**, you're choosing to scan all rows in the data source. This is a comprehensive scan to ensure that PII entities are identified.

When you choose **Detect fields containing PII**, you're choosing to scan a sample of rows for PII entities. This is a way to keep costs and resources low while also identifying the fields where PII entities are found.

When you choose to detect fields that contain PII, you can reduce costs and improve performance by sampling a portion of rows. Choosing this option will allow you to specify additional options:

- **Sample portion:** This allows you to specify the percentage of rows to sample. For example, if you enter '50', you're specifying that you want 50 percent of scanned rows for the PII entity.

- **Detection threshold:** This allows you to specify the percentage of rows that contain the PII entity in order for the entire column to be identified as having the PII entity. For example, if you enter ‘10’, you’re specifying that the number of the PII entity, US Phone, in the rows that are scanned must be 10 percent or greater in order for the field to be identified as having the PII entity, US Phone. If the percentage of rows that contain the PII entity is less than 10 percent, that field will not be labeled as having the PII entity, US Phone, in it.

Choosing the PII entities to detect

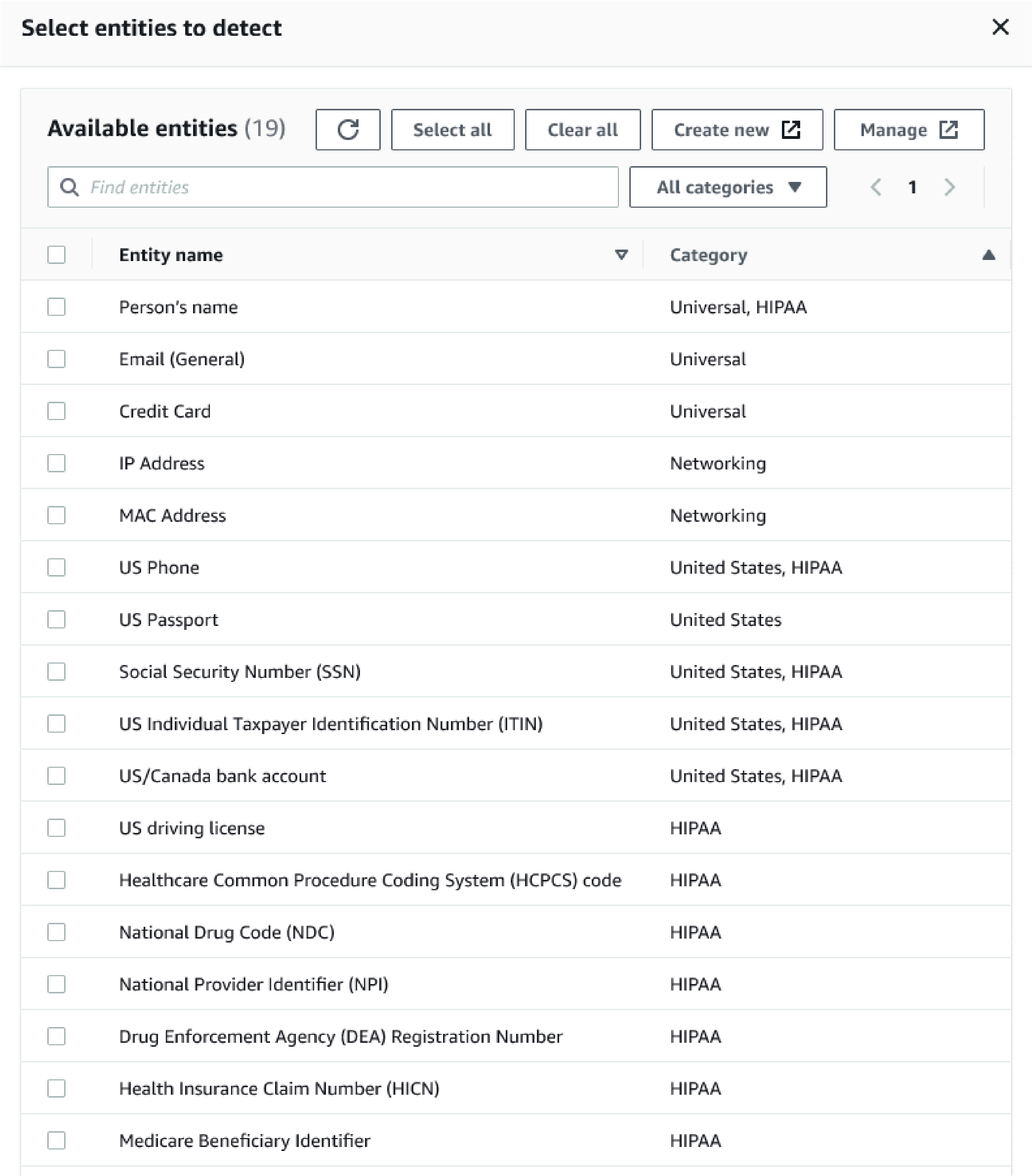
If you chose **Detect PII in each cell**, you can choose from one of three options:

- All available PII patterns - this includes AWS entities.
- Select categories - when you select categories, PII patterns will automatically include patterns in the categories that you select.
- Select specific patterns - Only the patterns that you select will be detected.

For a full list of managed sensitive data types, see Managed data types.

Choose from all available PII patterns

If you choose **All available PII patterns**, select entities pre-defined by AWS. You can select one, more than one, or all entities.



Select entities to detect

Available entities (19)

☐	Entity name	Category
☐	Person’s name	Universal, HIPAA
☐	Email (General)	Universal
☐	Credit Card	Universal
☐	IP Address	Networking
☐	MAC Address	Networking
☐	US Phone	United States, HIPAA
☐	US Passport	United States
☐	Social Security Number (SSN)	United States, HIPAA
☐	US Individual Taxpayer Identification Number (ITIN)	United States, HIPAA
☐	US/Canada bank account	United States, HIPAA
☐	US driving license	HIPAA
☐	Healthcare Common Procedure Coding System (HCPCS) code	HIPAA
☐	National Drug Code (NDC)	HIPAA
☐	National Provider Identifier (NPI)	HIPAA
☐	Drug Enforcement Agency (DEA) Registration Number	HIPAA
☐	Health Insurance Claim Number (HICN)	HIPAA
☐	Medicare Beneficiary Identifier	HIPAA

Select categories

If you chose **Select categories** as the PII patterns to detect, you can select from the options in the drop-down menu. Note that some entities can belong to more than one category. For example, *Person's name* is an entity that belongs to the *Universal* and *HIPAA* categories.

- Universal (examples: Email, Credit Card)
- HIPAA (examples: US Driving License, Healthcare Common Procedure Coding System (HCPCS) code)
- Networking (examples: IP Address, MAC Address)
- Argentina
- Australia
- Austria
- Belgium
- Bosnia
- Bulgaria
- Canada
- Chile
- Colombia
- Croatia
- Cyprus
- Czechia
- Denmark
- Estonia
- Finland
- France
- Germany
- Greece
- Hungary
- Ireland
- Korea
- Japan

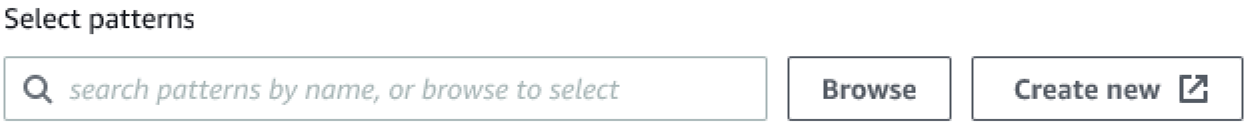
- Mexico
- Netherlands
- New Zealand
- Norway
- Portugal
- Romania
- Singapore
- Slovakia
- Slovenia
- Spain
- Sweden
- Switzerland
- Turkey
- Ukraine
- United States
- United Kingdom
- Venezuela

Select specific patterns

If you choose **Select specific patterns** as the PII patterns to detect, you can search or browse from a list of patterns you've already created, or create a new detection entity pattern.

The steps below describe how to create a new custom pattern for detecting sensitive data. You will create the custom pattern by entering a name for the custom pattern, add a regular expression, and optionally, define context words.

1. To create a new pattern, click the **Create new** button.



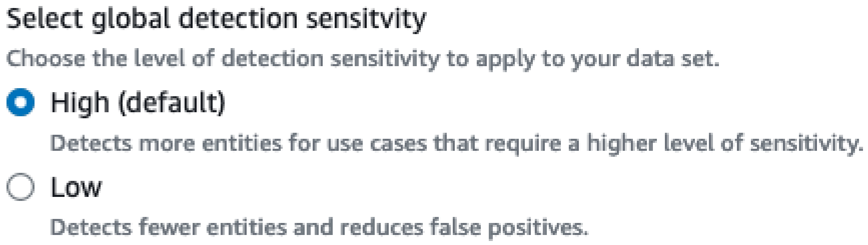
2. In the Create detection entity page, enter the entity name and a regular expression. The regular expression (Regex) is what AWS Glue will use to match entities.
3. Click **Validate**. If the validation is successful, you will see a confirmation message stating that the string is a valid regular expression. If the validation is not successful, you will see a message stating that the string does not conform to proper formatting and accepted character literals, operators or constructs.
4. You can choose to add Context words in addition to the regular expression. Context words may increase the likelihood of a match. These can be useful in cases where field names are not descriptive of the entity. For example, social security numbers may be named 'SSN' or 'SS'. Adding these context words can help match the entity.
5. Click **Create** to create the detection entity. Any created entities are visible in the AWS Glue Studio console. Click on **Detection entities** in the left-hand navigation menu.

   You can edit, delete, or create detection entities from the **Detection entities** page. You can also search for a pattern using the search field.

Specifying the level of detection sensitivity

You can set the level of sensitivity when using detecting sensitive data.

- **High –** (Default) Detects more entities for use cases that require a higher level of sensitivity. All AWS Glue jobs created after November 2023 are automatically opted-in to this setting.
- **Low –** Detects fewer entities and reduces false positives.



Select global detection sensitvity

Choose the level of detection sensitivity to apply to your data set.

● High (default)

Detects more entities for use cases that require a higher level of sensitivity.

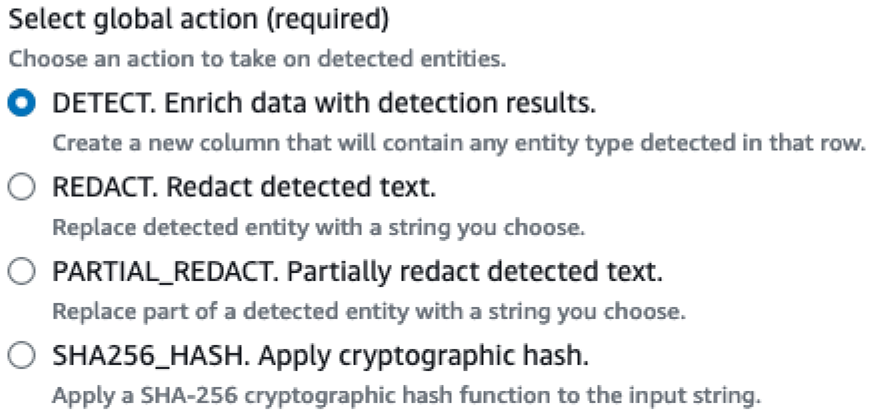
○ Low

Detects fewer entities and reduces false positives.

Choosing what to do with identified PII data

If you chose to detect PII in the entire data source, you can select a global action to apply:

- **Enrich data with detection results:** If you chose Detect PII in each cell, you can store the detected entities into a new column.
- **Redact detected text:** You can replace the detected PII value with a string that you specify in the optional Replacing text input field. If no string is specified, the detected PII entity is replaced with '*******'.
- **Partially redact detected text:** You can replace part of the detected PII value with a string you choose. There are two possible options: to either leave the ends unmasked or to mask by providing an explicit regex pattern. This feature is not available in AWS Glue 2.0.
- **Apply cryptographic hash:** You can pass the detected PII value to a SHA-256 cryptographic hash function and replace the value with the function’s output.



Select global action (required)

Choose an action to take on detected entities.

● DETECT. Enrich data with detection results.

Create a new column that will contain any entity type detected in that row.

○ REDACT. Redact detected text.

Replace detected entity with a string you choose.

○ PARTIAL_REDACT. Partially redact detected text.

Replace part of a detected entity with a string you choose.

○ SHA256_HASH. Apply cryptographic hash.

Apply a SHA-256 cryptographic hash function to the input string.

Differences between AWS Glue versions 2.0 and 3.0+

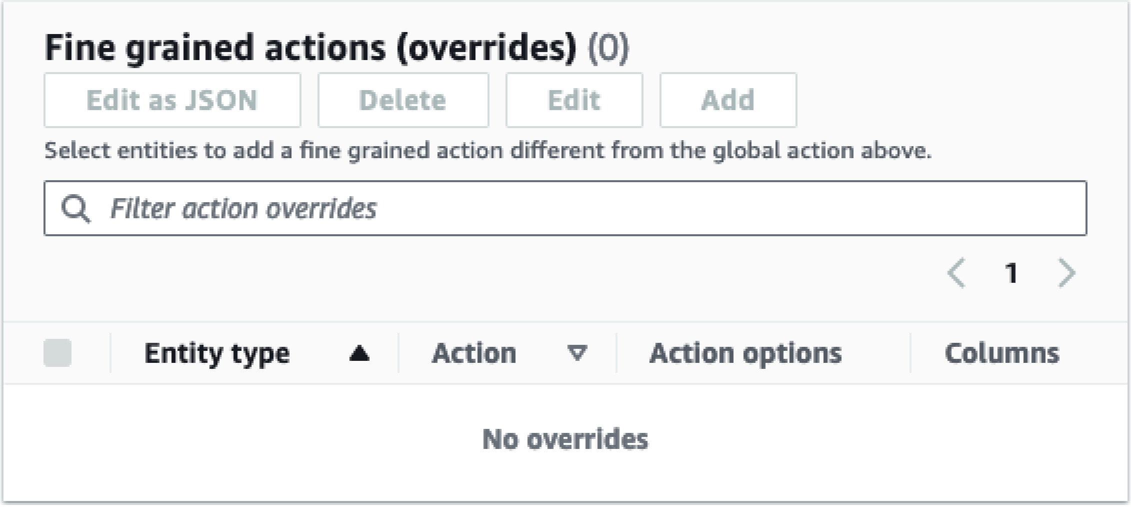
AWS Glue 2.0 jobs will return a new DataFrame with the detected PII information for each column in a supplementary column. Any redaction or hash work is visible within the AWS Glue script in the visual tab.

AWS Glue 3.0 and 4.0 jobs will return a new DataFrame with this same supplementary column. A new key for “actionUsed” is present and can be one of `DETECT`, `REDACT`, `PARTIAL_REDACT`, or `SHA256_HASH`. If a masking action is selected, the DataFrame will return data with sensitive data masked.

Adding fine-grained action overrides

Additional detection and action settings can be added to the fine-grained actions overrides table. This allows you to:

- **Include or exclude certain columns from detection** – An inferred schema on the data source will populate the table with available columns.
- **Specify specific settings that are more fine-grained than using global actions** – For example, you can specify different redaction text settings for different entity types.
- **Specify a different action than the global action** – If a different action wants to be applied on a different sensitive data type, that can be done here. Note that two different edit-in-place actions (redaction and hashing) cannot be used on the same column, but detect can always be used.


Fine grained actions (overrides) (0)

Select entities to add a fine grained action different from the global action above.

☐	Entity type ▲	Action ▽	Action options	Columns

No overrides

Managing ETL jobs with AWS Glue Studio

You can use the simple graphical interface in AWS Glue Studio to manage your ETL jobs. Using the navigation menu, choose **Jobs** to view the **Jobs** page. On this page, you can see all the jobs that you have created either with AWS Glue Studio or the AWS Glue console. You can view, manage, and run your jobs on this page.

You can also perform the following tasks on this page:

- Start a job run
- Schedule job runs
- Manage job schedules
- Stop job runs
- View your jobs
- View information for recent job runs

- View the job script
- Modify the job properties
- Save the job
- Clone a job
- Delete jobs

Start a job run

In AWS Glue Studio, you can run your jobs on demand. A job can run multiple times, and each time you run the job, AWS Glue collects information about the job activities and performance. This information is referred to as a *job run* and is identified by a job run ID.

You can initiate a job run in the following ways in AWS Glue Studio:

- On the **Jobs** page, choose the job you want to start, and then choose the **Run job** button.
- If you're viewing a job in the visual editor and the job has been saved, you can choose the **Run** button to start a job run.

For more information about job runs, see Working with Jobs on the AWS Glue Console in the *AWS Glue Developer Guide*.

Schedule job runs

In AWS Glue Studio, you can create a schedule to have your jobs run at specific times. You can specify constraints, such as the number of times that the jobs run, which days of the week they run, and at what time. These constraints are based on `cron` and have the same limitations as `cron`. For example, if you choose to run your job on day 31 of each month, keep in mind that some months don't have 31 days. For more information about `cron`, see Cron Expressions in the *AWS Glue Developer Guide*.

To run jobs according to a schedule

1. Create a job schedule using one of the following methods:
    - On the **Jobs** page, choose the job you want to create a schedule for, choose **Actions**, and then choose **Schedule job**.

  - If you're viewing a job in the visual editor and the job has been saved, choose the **Schedules** tab. Then choose **Create Schedule**.

2. On the **Schedule job run** page, enter the following information:

   - **Name**: Enter a name for your job schedule.
   - **Frequency**: Enter the frequency for the job schedule. You can choose the following:
     - **Hourly**: The job will run every hour, starting at a specific minute. You can specify the **Minute** of the hour that the job should run. By default, when you choose hourly, the job runs at the beginning of the hour (minute 0).
     - **Daily**: The job will run every day, starting at a time. You can specify the **Minute** of the hour that the job should run and the **Start hour** for the job. Hours are specified using a 23-hour clock, where you use the numbers 13 to 23 for the afternoon hours. The default values for minute and hour are 0, which means that if you select **Daily**, the job by default will run at midnight.
     - **Weekly**: The job will run every week on one or more days. In addition to the same settings described previous for Daily, you can choose the days of the week on which the job will run. You can choose one or more days.
     - **Monthly**: The job will run every month on a specific day. In addition to the same settings described previous for Daily, you can choose the day of the month on which the job will run. Specify the day as a numeric value from 1 to 31. If you select a day that does not exist in a month, for example the 30th day of February, then the job does not run that month.
     - **Custom**: Enter an expression for your job schedule using the `cron` syntax. Cron expressions allow you to create more complicated schedules, such as the last day of the month (instead of a specific day of the month) or every third month on the 7th and 21st days of the month.

       See Cron Expressions in the *AWS Glue Developer Guide*
   - **Description**: You can optionally enter a description for your job schedule. If you plan to use the same schedule for multiple jobs, having a description makes it easier to determine what the job schedule does.

3. Choose **Create schedule** to save the job schedule.

4. After you create the schedule, a success message appears at the top of the console page. You can choose **Job details** in this banner to view the job details. This opens the visual job editor page, with the **Schedules** tab selected.

Manage job schedules

After you have created schedules for a job, you can open the job in the visual editor and choose the **Schedules** tab to manage the schedules.

On the **Schedules** tab of the visual editor, you can perform the following tasks:

- Create a new schedule.

  Choose **Create schedule**, then enter the information for your schedule as described in the section called "Schedule job runs".
- Edit an existing schedule.

  Choose the schedule you want to edit, then choose **Action** followed by **Edit schedule**. When you choose to edit an existing schedule, the **Frequency** shows as **Custom**, and the schedule is displayed as a `cron` expression. You can either modify the `cron` expression, or specify a new schedule using the **Frequency** button. When you finish with your changes, choose **Update schedule**.
- Pause an active schedule.

  Choose an active schedule, and then choose **Action** followed by **Pause schedule**. The schedule is instantly deactivated. Choose the refresh (reload) button to see the updated job schedule status.
- Resume a paused schedule.

  Choose a deactivated schedule, and then choose **Action** followed by **Resume schedule**. The schedule is instantly activated. Choose the refresh (reload) button to see the updated job schedule status.
- Delete a schedule.

  Choose the schedule you want to remove, and then choose **Action** followed by **Delete schedule**. The schedule is instantly deleted. Choose the refresh (reload) button to see the updated job schedule list. The schedule will show a status of **Deleting** until it has been completely removed.

Stop job runs

You can stop a job before it has completed its job run. You might choose this option if you know that the job isn't configured correctly, or if the job is taking too long to complete.

On the **Monitoring** page, in the **Job runs** list, choose the job that you want to stop, choose **Actions**, and then choose **Stop run**.

View your jobs

You can view all your jobs on the **Jobs** page. You can access this page by choosing **Jobs** in the navigation pane.

On the **Jobs** page, you can see all the jobs that were created in your account. The **Your jobs** list shows the job name, its type, the status of the last run of that job, and the dates on which the job was created and last modified. You can choose the name of a job to see detailed information for that job.

You can also use the Monitoring dashboard to view all your jobs. You can access the dashboard by choosing **Monitoring** in the navigation pane.

Customize the job display

You can customize how the jobs are displayed in the **Your jobs** section of the **Jobs** page. Also, you can enter text in the search text field to display only jobs with a name that contains that text.

If you choose the settings icon

⚙

in the **Your jobs** section, you can customize how AWS Glue Studio displays the information in the table. You can choose to wrap the lines of text in the display, change the number of jobs displayed on the page, and specify which columns to display.

View information for recent job runs

A job can run multiple times as new data is added at the source location. Each time a job runs, the job run is assigned a unique ID, and information about that job run is collected. You can view this information by using the following methods:

- Choose the **Runs** tab of the visual editor to view the job run information for the currently displayed job.

  On the **Runs** tab (the **Recent job runs** page), there is a card for each job run. The information displayed on the **Runs** tab includes:
  - Job run ID

  - Number of attempts to run this job
  - Status of the job run
  - Start and end time for the job run
  - The runtime for the job run
  - A link to the job log files
  - A link to the job error log files
  - The error returned for failed jobs
- You can select a job run to view additional information about the job, including the following:
  - **Input arguments**
  - **Continuous logs**
  - **Metrics** – You can see visualizations of basic metrics. For more information on included metrics, see the section called "Viewing Amazon CloudWatch metrics for a Spark job run".
  - **Spark UI** – You can visualize Spark logs for your job in the Spark UI. For more information about using the Spark Web UI, see the section called "Monitoring with the Spark UI". Enable this feature by following the procedure in the section called "Enabling the Spark UI for jobs".

You can select **View details** to view similar information on the job run details page. Alternatively, you can navigate to the job run details page through the **Monitoring** page. In the navigation pane, choose **Monitoring**. Scroll down to the **Job runs** list. Choose the job and then choose **View run details**. The contents are described in Viewing the details of a job run.

For more information about the job logs, see Viewing the job run logs.

View the job script

After you provide information for all the nodes in the job, AWS Glue Studio generates a script that is used by the job to read the data from the source, transform the data, and write the data in the target location. If you save the job, you can view this script at any time.

To view the generated script for your job

1. Choose **Jobs** in the navigation pane.
2. On the **Jobs** page, in the **Your Jobs** list, choose the name of the job you want to review. Alternatively, you can select a job in the list, choose the **Actions** menu, and then choose **Edit job**.

3. On the visual editor page, choose the **Script** tab at the top to view the job script.

   If you want to edit the job script, see AWS Glue programming guide.

Modify the job properties

The nodes in the job diagram define the actions performed by the job, but there are several properties that you can configure for the job as well. These properties determine the environment that the job runs in, the resources it uses, the threshold settings, the security settings, and more.

To customize the job run environment

1. Choose **Jobs** in the navigation pane.
2. On the **Jobs** page, in the **Your Jobs** list, choose the name of the job you want to review.
3. On the visual editor page, choose the **Job details** tab at the top of the job editing pane.
4. Modify the job properties, as needed.

   For more information about the job properties, see Defining Job Properties in the *AWS Glue Developer Guide*.
5. Expand the **Advanced properties** section if you need to specify these additional job properties:

   - **Script filename** – The name of the file that stores the job script in Amazon S3.
   - **Script path** – The Amazon S3 location where the job script is stored.
   - **Job metrics** – (Not available for Python shell jobs) Turns on the creation of Amazon CloudWatch metrics when this job runs.
   - **Continuous logging** – (Not available for Python shell jobs) Turns on continuous logging to CloudWatch, so the logs are available for viewing before the job completes.
   - **Spark UI** and **Spark UI logs path** – (Not available for Python shell jobs) Turns on the use of Spark UI for monitoring this job and specifies the location for the Spark UI logs.
   - **Maximum concurrency** – Sets the maximum number of concurrent runs that are allowed for this job.
   - **Temporary path** – The location of a working directory in Amazon S3 where temporary intermediate results are written when AWS Glue runs the job script.
   - **Delay notification threshold (minutes)** – Specify a delay threshold for the job. If the job runs for a longer time than that specified by the threshold, then AWS Glue sends a delay notification for the job to CloudWatch.

- **Security configuration** and **Server-side encryption** – Use these fields to choose the encryption options for the job.
- **Use Glue Data Catalog as the Hive metastore** – Choose this option if you want to use the AWS Glue Data Catalog as an alternative to Apache Hive Metastore.
- **Additional network connection** – For a data source in a VPC, you can specify a connection of type `Network` to ensure your job accesses your data through the VPC.
- **Python library path**, **Dependent jars path** (Not available for Python shell jobs), or **Referenced files path** – Use these fields to specify the location of additional files used by the job when it runs the script.
- **Job Parameters** – You can add a set of key-value pairs that are passed as named parameters to the job script. In Python calls to AWS Glue APIs, it's best to pass parameters explicitly by name. For more information about using parameters in a job script, see Passing and Accessing Python Parameters in AWS Glue in the *AWS Glue Developer Guide*.
- **Tags** – You can add tags to the job to help you organize and identify them.

6. After you have modified the job properties, save the job.

Store Spark shuffle files on Amazon S3

Some ETL jobs require reading and combining information from multiple partitions, for example, when using a join transform. This operation is referred to as *shuffling*. During a shuffle, data is written to disk and transferred across the network. With AWS Glue version 3.0, you can configure Amazon S3 as a storage location for these files. AWS Glue provides a shuffle manager which writes and reads shuffle files to and from Amazon S3. Writing and reading shuffle files from Amazon S3 is slower (by 5%-20%) compared to local disk (or Amazon EBS which is heavily optimized for Amazon EC2). However, Amazon S3 provides unlimited storage capacity, so you don't have to worry about "`No space left on device`" errors when running your job.

To configure your job to use Amazon S3 for shuffle files

1. On the **Jobs** page, in the **Your Jobs** list, choose the name of the job you want to modify.
2. On the visual editor page, choose the **Job details** tab at the top of the job editing pane.

   Scroll down to the **Job parameters** section.
3. Specify the following key-value pairs.
   - `--write-shuffle-files-to-s3` — `true`

This is the main parameter that configures the shuffle manager in AWS Glue to use Amazon S3 buckets for writing and reading shuffle data. By default, this parameter has a value of `false`.

- (Optional) `--write-shuffle-spills-to-s3` — `true`

  This parameter allows you to offload spill files to Amazon S3 buckets, which provides additional resiliency to your Spark job in AWS Glue. This is only required for large workloads that spill a lot of data to disk. By default, this parameter has a value of `false`.

- (Optional) `--conf spark.shuffle.glue.s3ShuffleBucket` — `S3://<shuffle-bucket>`

  This parameter specifies the Amazon S3 bucket to use when writing the shuffle files. If you do not set this parameter, the location is the `shuffle-data` folder in the location specified for **Temporary path** (`--TempDir`).

  > **Note**
  >
  > Make sure the location of the shuffle bucket is in the same AWS Region in which the job runs.
  > Also, the shuffle service does not clean the files after the job finishes running, so you should configure the Amazon S3 storage life cycle policies on the shuffle bucket location. For more information, see Managing your storage lifecycle in the *Amazon S3 User Guide*.

Save the job

A red **Job has not been saved** callout is displayed to the left of the **Save** button until you save the job.

Job has not been saved
Save

To save your job

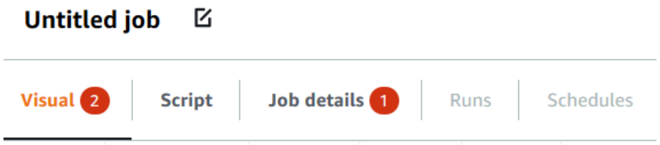
1. Provide all the required information in the **Visual** and **Job details** tabs.
2. Choose the **Save** button.

After you save the job, the 'not saved' callout changes to display the time and date that the job was last saved.

If you exit AWS Glue Studio before saving your job, the next time you sign in to AWS Glue Studio, a notification appears. The notification indicates that there is an unsaved job, and asks if you want to restore it. If you choose to restore the job, you can continue to edit it.

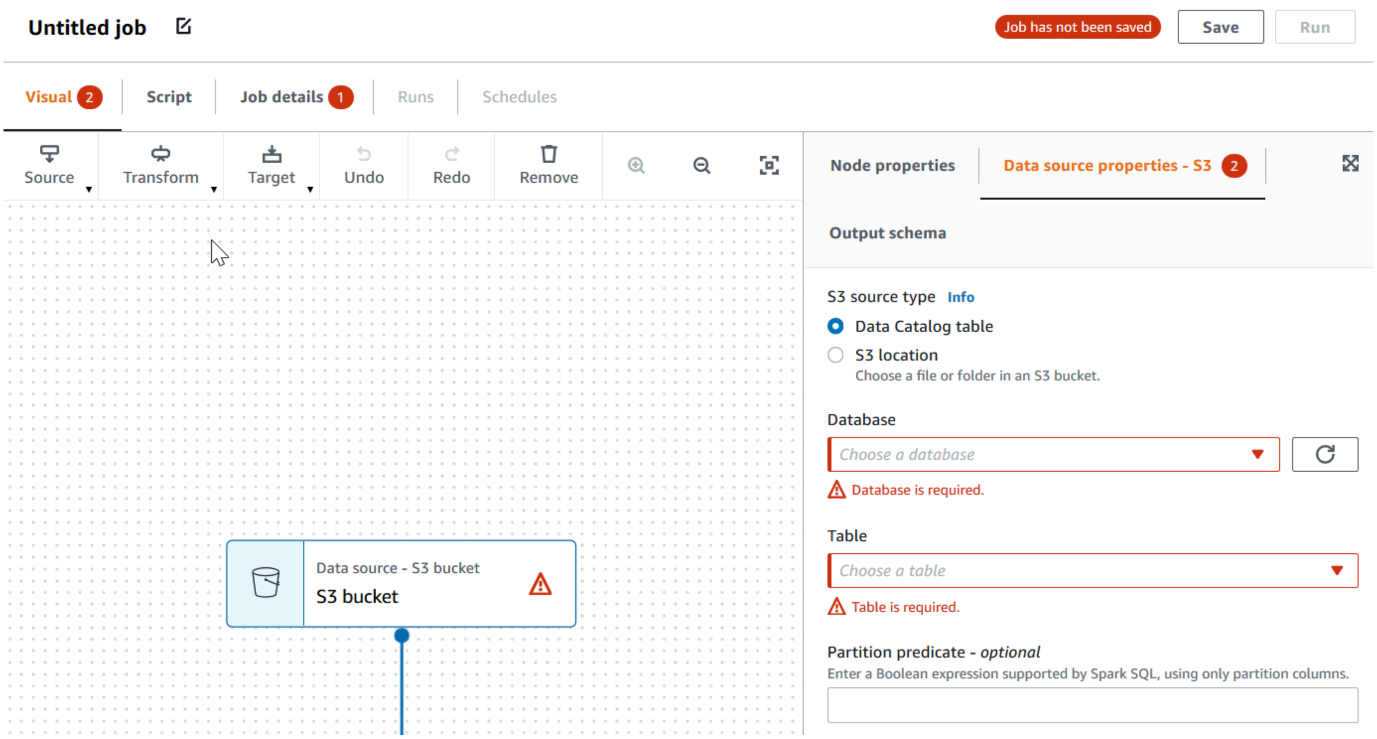
Troubleshooting errors when saving a job

If you choose the **Save** button, but your job is missing some required information, then a red callout appears on the tab where the information is missing. The number in the callout indicates how many missing fields were detected.



- If a node in the visual editor isn't configured correctly, the **Visual** tab shows a red callout, and the node with the error displays a warning symbol ⚠ .

  1. Choose the node. In the node details panel, a red callout appears on the tab where the missing or incorrect information is located.
  2. Choose the tab in the node details panel that shows a red callout, and then locate the problem fields, which are highlighted. An error message below the fields provides additional information about the problem.


Untitled job
Job has not been saved
Save
Run
Visual 2
Script
Job details 1
Runs
Schedules
Source
Transform
Target
Undo
Redo
Remove
Node properties
Data source properties - S3 2
Output schema
Data source - S3 bucket
S3 bucket

S3 source type Info

● Data Catalog table

○ S3 location

Choose a file or folder in an S3 bucket.

Database

Choose a database

Database is required.

Table

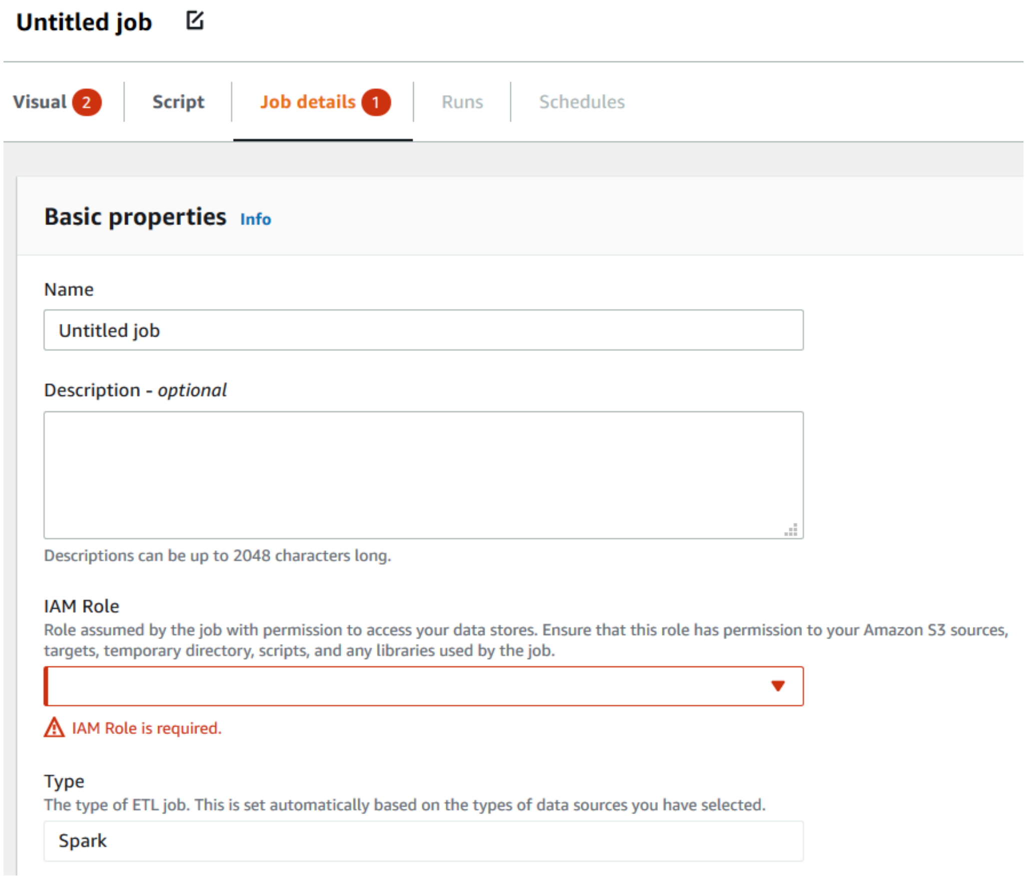
Choose a table

Table is required.

Partition predicate - *optional*

Enter a Boolean expression supported by Spark SQL, using only partition columns.

- If there is a problem with the job properties, the **Job details** tab shows a red callout. Choose that tab and locate the problem fields, which are highlighted. The error messages below the fields provide additional information about the problem.



Name

Untitled job

Description - *optional*

Descriptions can be up to 2048 characters long.

IAM Role

Role assumed by the job with permission to access your data stores. Ensure that this role has permission to your Amazon S3 sources, targets, temporary directory, scripts, and any libraries used by the job.

IAM Role is required.

Type

The type of ETL job. This is set automatically based on the types of data sources you have selected.

Spark

Clone a job

You can use the **Clone job** action to copy an existing job into a new job.

To create a new job by copying an existing job

1. On the **Jobs** page, in the **Your jobs** list, choose the job that you want to duplicate.
2. From the **Actions** menu, choose **Clone job**.
3. Enter a name for the new job. You can then save or edit the job.

Delete jobs

You can remove jobs that are no longer needed. You can delete one or more jobs in a single operation.

To remove jobs from AWS Glue Studio

1. On the **Jobs** page, in the **Your jobs** list, choose the jobs that you want to delete.
2. From the **Actions** menu, choose **Delete job**.
3. Verify that you want to delete the job by entering **`delete`**.

You can also delete a saved job when you're viewing the **Job details** tab for that job in the visual editor.

Working with jobs in AWS Glue

The following sections provide information on ETL and Ray jobs in AWS Glue.

Topics

- AWS Glue versions
- Working with Spark jobs in AWS Glue
- Working with Ray jobs in AWS Glue
- Configuring job properties for Python shell jobs in AWS Glue
- Monitoring AWS Glue

AWS Glue versions

You can configure the AWS Glue version parameter when you add or update a job. The AWS Glue version determines the versions of Apache Spark and Python that AWS Glue supports. The Python version indicates the version that's supported for jobs of type Spark. The following table lists the available AWS Glue versions, the corresponding Spark and Python versions, and other changes in functionality.

AWS Glue versions

AWS Glue version	Supported runtime environment versions	Supported Java version	Changes in functionality
AWS Glue 5.0	• Spark 3.5.4 • Python 3.11 • Scala 2.12.18	Java 17	In addition to the framework updates, there are optimizations and upgrades built into this AWS Glue release, such as: • Amazon SageMaker Unified Studio support

AWS Glue version	Supported runtime environment versions	Supported Java version	Changes in functionality
			• Amazon SageMaker Lakehouse support • Open Table Formats (OTF) updated to Hudi 0.15.0, Iceberg 1.7.1, and Delta Lake 3.3.0 • Spark-native fine-grained access control using Lake Formation. • Amazon S3 Access Grants support • `requirements.txt` support to install additional Python libraries • Data lineage support in Amazon DataZone • Amazon S3 Table Bucket support • AWS Glue Data Catalog multi-dialect view support **Limitations**

AWS Glue version	Supported runtime environment versions	Supported Java version	Changes in functionality
			The following are limitations with AWS Glue 5.0: - Glue Dynamic Frame / `GlueContext` - based table-level access control with AWS Lake Formation permissions supported in Glue 4.0 or before is not supported in Glue 5.0. Use the new Spark native fine-grained access control (FGAC) in Glue 5.0. For more information about migrating to AWS Glue version 5.0, see Migrating AWS Glue for Spark jobs to AWS Glue version 5.0.

AWS Glue version	Supported runtime environment versions	Supported Java version	Changes in functionality

AWS Glue 4.0	**Spark environment versions**
	- Spark 3.3.0
	- Python 3.10
		Java 8	AWS Glue 4.0 has a number of optimizations and upgrades built into this AWS Glue release, such as:
			- Many Spark functionality upgrades from Spark 3.1 to Spark 3.3:
			  - Several functionality improvements when paired with Pandas. For more information, see [What's New in Spark 3.3](#).
			  - Additional optimizations developed on Amazon EMR.
			  - Upgrade to EMR File System (EMRFS) 2.53.
			- Log4j 2 migration from Log4j 1.x
			- Several Python module updates from AWS Glue 3.0, such as an

AWS Glue version	Supported runtime environment versions	Supported Java version	Changes in functionality
			upgraded version of Boto. • Upgrade of several connectors, including the default Amazon Redshift connector. See [Appendix C: Connector upgrades](#). • Upgrade of several JDBC drivers. See [Appendix B: JDBC driver upgrades](#). • Updated with a new Amazon Redshift connector and JDBC driver. • Native support for open-data lake frameworks with Apache Hudi, Delta Lake, and Apache Iceberg. • Native support for the Amazon S3-based Cloud Shuffle Storage Plugin (an Apache Spark plugin) to use Amazon S3 for shuffling and

AWS Glue version	Supported runtime environment versions	Supported Java version	Changes in functionality
			elastic storage capacity. **Limitations** The following are limitations with AWS Glue 4.0: - AWS Glue machine learning and personally identifiable information (PII) transforms are not yet available in AWS Glue 4.0. For more information about migrating to AWS Glue version 4.0, see Migrating AWS Glue for Spark jobs to AWS Glue version 4.0.

AWS Glue version	Supported runtime environment versions	Supported Java version	Changes in functionality
	Ray environment versions - Ray 2.4.0 Python 3.9	N/A	Build and run distributed Python applications with AWS Glue for Ray. - Supports Ray-2.4.0 data distribution (`ray[data]`) with Python 3.9. For more information on this Ray release, see Ray-2.4.0 in the Ray GitHub repository. - Supports installing additional Python libraries into the `Ray2.4` runtime environment. For more information, see the section called "Additional Python modules for Ray jobs". - Integrates logs and metrics from Ray jobs with Amazon CloudWatch. For more information, see the section called "Troubleshooting Ray errors"

AWS Glue version	Supported runtime environment versions	Supported Java version	Changes in functionality
			and [the section called "Ray job metrics"](#). - Aggregates and visualizes metrics for Ray jobs in AWS Glue Studio, on each job run page. - Supports distributing files to each working directory across your cluster, spilling objects from the Ray object store to Amazon S3, and controlling the minimum number of worker nodes allocated to your Ray job. For more information, see [the section called "Ray job parameters"](#). **Limitations on Ray jobs in AWS Glue 4.0** - AWS Glue interactive sessions for Ray remain in preview for this release.

AWS Glue version	Supported runtime environment versions	Supported Java version	Changes in functionality
			• AWS Glue for Ray integration with Amazon VPC is not currently available . Resources in a VPC in AWS will not be accessible without a public route. For more information about using AWS Glue with Amazon VPC, see the section called "Configuring interface VPC endpoints (AWS PrivateLink) for AWS Glue ". • AWS Glue for Ray is available in US East (N. Virginia), US East (Ohio), US West (Oregon), Asia Pacific (Tokyo), and Europe (Ireland).

AWS Glue version	Supported runtime environment versions	Supported Java version	Changes in functionality

AWS Glue 3.0	- Spark 3.1.1
	- Python 3.7
		Java 8	In addition to the Spark engine upgrade to 3.0, there are optimizations and upgrades built into this AWS Glue release, such as:
			- Builds the AWS Glue ETL Library against Spark 3.0, which is a major release for Spark.
			- Streaming jobs are supported on AWS Glue 3.0.
			- Includes new AWS Glue Spark runtime optimizations for performance and reliability:
			  - Faster in-memory columnar processing based on Apache Arrow for reading CSV data.
			  - SIMD-based execution for vectorized reads with CSV data.

AWS Glue version	Supported runtime environment versions	Supported Java version	Changes in functionality
			- Spark upgrade also includes additional optimizations developed on Amazon EMR. - Upgraded EMRFS from 2.38 to 2.46 enabling new features and bug fixes for Amazon S3 access. - Upgraded several dependencies that were required for the new Spark version. - Upgraded JDBC drivers for our natively supported data sources. **Limitations** The following are limitations with AWS Glue 3.0: - AWS Glue machine learning transforms are not yet

AWS Glue version	Supported runtime environment versions	Supported Java version	Changes in functionality
			available in AWS Glue 3.0. • Some custom Spark connectors do not work with AWS Glue 3.0 if they depend on Spark 2.4 and do not have compatibility with Spark 3.1.

AWS Glue version	Supported runtime environment versions	Supported Java version	Changes in functionality
AWS Glue 2.0 ([end of life on April 1, 2026]())	• Spark 2.4.3 • Python 3.7	N/A	In addition to the features provided in AWS Glue version 1.0, AWS Glue version 2.0 also provides: • An upgraded infrastructure for running Apache Spark ETL jobs in AWS Glue with reduced startup times. • Default logging is now real time, with separate streams for drivers and executors, and outputs and errors. • Support for specifying additional Python modules or different versions at the job level. **Note** AWS Glue version 2.0 differs from AWS Glue

AWS Glue version	Supported runtime environment versions	Supported Java version	Changes in functionality
			version 1.0 for some dependencies and versions due to underlying architectural changes. Validate your AWS Glue jobs before migrating across major AWS Glue version releases.

AWS Glue version	Supported runtime environment versions	Supported Java version	Changes in functionality
AWS Glue 1.0 (end of life on April 1, 2026)	• Spark 2.4.3 • Python 2.7 • Python 3.6	N/A	You can maintain job bookmarks for Parquet and ORC formats in AWS Glue ETL jobs (using AWS Glue version 1.0). Previously, you were only able to bookmark common Amazon S3 source formats such as JSON, CSV, Apache Avro, and XML in AWS Glue ETL jobs. When setting format options for ETL inputs and outputs, you can specify to use Apache Avro reader/writer format 1.8 to support Avro logical type reading and writing (using AWS Glue version 1.0). Previously, only the version 1.7 Avro reader/writer format was supported. The DynamoDB connection type supports a writer

AWS Glue version	Supported runtime environment versions	Supported Java version	Changes in functionality
			option (using AWS Glue version 1.0). **Limitations** The following are limitations with AWS Glue 1.0: - AWS Glue versions 0.9 and 1.0 are not available in the Asia Pacific (Jakarta) (`ap-southeast-3`), Middle East (UAE) (`me-central-1`), or other new Regions going forward.

AWS Glue version	Supported runtime environment versions	Supported Java version	Changes in functionality
AWS Glue 0.9 (end of life on April 1, 2026)	Spark 2.2.1 Python 2.7	N/A	Jobs that were created without specifying an AWS Glue version default to AWS Glue 0.9. **Limitations** The following are limitations with AWS Glue 0.9: - AWS Glue versions 0.9 and 1.0 are not available in the Asia Pacific (Jakarta) (`ap-southeast-3`), Middle East (UAE) (`me-central-1`), or other new Regions going forward.

> **Note**
>
> The following Glue versions support these versions of PythonShell:
>
> - PythonShell v3.6 is supported in Glue version 1.0.
> - PythonShell v3.9 is supported in Glue version 3.0.
>
> Additionally, dev endpoints are supported only in Glue version 1.0, and 0.9.

AWS Glue version support policy

AWS Glue is a serverless data integration service that makes it easy to discover, prepare, and combine data for analytics, machine learning, and application development. An *AWS Glue job* contains the business logic that performs the data integration work in AWS Glue. There are three types of jobs in AWS Glue: *Spark (batch and streaming)*, *Ray* and *Python shell*. When you define your job, you specify the AWS Glue version, which configures versions in the underlying Spark, Ray or Python runtime environment. For example: an AWS Glue version 5.0 Spark job supports Spark 3.5.4 and Python 3.11.

Support policy

AWS Glue versions are built around a combination of operating system, programming language, and software libraries that are subject to maintenance and security updates. AWS Glue's version support policy is to end support for a version when any major component of the version reaches the end of community long-term support (LTS) and security updates are no longer available. AWS Glue's version support policy includes the following statuses:

End of Support (EOS) - When an AWS Glue version reaches EOS:

- AWS Glue will no longer apply security patches or other updates to EOS versions.
- AWS Glue jobs on EOS versions are not eligible for technical support.
- AWS Glue may not honor SLAs when jobs are run on EOS versions.

End of Life (EOL) - When an AWS Glue version reaches EOL:

- You can no longer create new AWS Glue jobs or interactive sessions on EOL versions.
- You can no longer start job runs on these AWS Glue versions.
- AWS Glue will stop existing job runs and interactive sessions on EOL versions.
- EOL versions will be removed from AWS Glue SDKs and APIs.

The following AWS Glue versions have reached end of support and will no longer be available after the end of life date. Changes to a version’s support status start at midnight (Pacific time zone) on the specified date.

Type	Glue version	End of support	End of life
Spark	Glue version 0.9 (Spark 2.2, Scala 2, Python 2)	6/1/2022	4/1/2026
Spark	Glue version 1.0 (Spark 2.4, Python 2)	6/1/2022	4/1/2026
Spark	Glue version 1.0 (Spark 2.4, Scala 2, Python 3)	9/30/2022	4/1/2026
Spark	Glue version 2.0 (Spark 2.4, Python 3)	1/31/2024	4/1/2026
Type	**Python version**	**End of support**	**End of life**
Python shell	Python 2 (AWS Glue Version 1.0)	6/1/2022	4/1/2026
Python shell	PythonShell 3.6 (Glue version 1.0)	3/31/2026	NA
Type	**Notebook version**	**End of support**	**End of life**
Development endpoint	Zeppelin notebook	9/30/2022	NA

Note

Creating new AWS Glue Python Shell 3.6 jobs will not be allowed once end of support is reached on March 31, 2026, but you can continue to update and run existing jobs. However, jobs running on discontinued versions are not eligible for technical support. AWS Glue will not apply security patches or other updates to discontinued versions. AWS Glue will also not honor SLAs when jobs are run on discontinued versions.

AWS strongly recommends that you migrate your jobs to supported versions.

For information on migrating your Spark jobs to the latest AWS Glue version, see Migrating AWS Glue jobs to AWS Glue version 5.0.

For migrating your Python shell jobs to the latest AWS Glue version:

- In the console, choose `Python 3 (Glue Version 4.0)`.
- In the CreateJob/UpdateJob API, set the `GlueVersion` parameter to `2.0`, and the `PythonVersion` to `3` under the `Command` parameter. The `GlueVersion` configuration does not affect the behavior of Python shell jobs, so there is no advantage to incrementing `GlueVersion`.
- You need to make your job script compatible with Python 3.

Migrating AWS Glue for Spark jobs to AWS Glue version 5.0

This topic describes the changes between AWS Glue versions 0.9, 1.0, 2.0, 3.0, and 4.0 to allow you to migrate your Spark applications and ETL jobs to AWS Glue 5.0. It also describes the features in AWS Glue 5.0 and the advantages of using it.

To use this feature with your AWS Glue ETL jobs, choose **`5.0`** for the `Glue version` when creating your jobs.

Topics

- New features
- Actions to migrate to AWS Glue 5.0
- Migration checklist
- AWS Glue 5.0 features
- Migrating from AWS Glue 4.0 to AWS Glue 5.0
- Migrating from AWS Glue 3.0 to AWS Glue 5.0
- Migrating from AWS Glue 2.0 to AWS Glue 5.0
- Logging behavior changes in AWS Glue 5.0
- Connector and JDBC driver migration for AWS Glue 5.0

New features

This section describes new features and advantages of AWS Glue version 5.0.

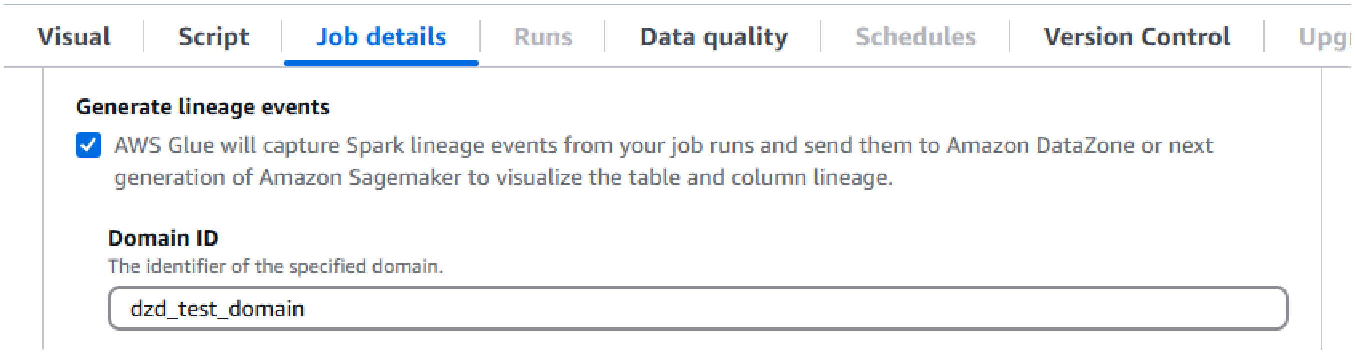
- Apache Spark update from 3.3.0 in AWS Glue 4.0 to 3.5.4 in AWS Glue 5.0. See Major enhancements from Spark 3.3.0 to Spark 3.5.4.
- Spark-native fine-grained access control (FGAC) using Lake Formation. This includes FGAC for Iceberg, Delta and Hudi tables. For more information, see Using AWS Glue with AWS Lake Formation for fine-grained access control.

  Note the following considerations or limitations for Spark-native FGAC:

  - Currently data writes are not supported
  - Writing into Iceberg through `GlueContext` using Lake Formation requires use of IAM access control instead

  For a complete list of limitations and considerations when using Spark-native FGAC, see the section called “Considerations”.
- Support for Amazon S3 Access Grants as a scalable access control solution to your Amazon S3 data from AWS Glue. For more information, see Using Amazon S3 Access Grants with AWS Glue.
- Open Table Formats (OTF) updated to Hudi 0.15.0, Iceberg 1.7.1, and Delta Lake 3.3.0
- Amazon SageMaker Unified Studio support.
- Amazon SageMaker Lakehouse and data abstraction integration. For more information, see Querying metastore data catalogs from AWS Glue ETL.
- Support to install additional Python libraries using `requirements.txt`. For more information, see Installing additional Python libraries in AWS Glue 5.0 or above using requirements.txt.
- AWS Glue 5.0 supports data lineage in Amazon DataZone. You can configure AWS Glue to automatically collect lineage information during Spark job runs and send the lineage events to be visualized in Amazon DataZone. For more information, see Data lineage in Amazon DataZone.

  To configure this on the AWS Glue console, turn on **Generate lineage events**, and enter your Amazon DataZone domain ID on the **Job details** tab.



Generate lineage events

☑ AWS Glue will capture Spark lineage events from your job runs and send them to Amazon DataZone or next generation of Amazon Sagemaker to visualize the table and column lineage.

Domain ID

The identifier of the specified domain.

dzd_test_domain

Alternatively, you can provide the following job parameter (provide your DataZone domain ID):

- Key: `--conf`
- Value:

```
extraListeners=io.openlineage.spark.agent.OpenLineageSparkListener
—conf spark.openlineage.transport.type=amazon_datazone_api
-conf spark.openlineage.transport.domainId=<your-domain-ID>
```

- Connector and JDBC driver updates. For more information, see Appendix B: JDBC driver upgrades and Appendix C: Connector upgrades.
- Java update from 8 to 17.
- Increased storage for AWS Glue `G.1X` and `G.2X` workers with disk space increasing to 94GB and 138GB respectively. Additionally, new worker types `G.12X`, `G.16X`, and memory-optimized `R.1X`, `R.2X`, `R.4X`, `R.8X` are available in AWS Glue 4.0 and later versions. For more information, see Jobs
- **Support for AWS SDK for Java, version 2** - AWS Glue 5.0 jobs can use the for Java versions 1.12.569 or 2.28.8 if the job supports v2. The AWS SDK for Java 2.x is a major rewrite of the version 1.x code base. It's built on top of Java 8+ and adds several frequently requested features. These include support for non-blocking I/O, and the ability to plug in a different HTTP implementation at runtime. For more information, including a Migration Guide from SDK for Java v1 to v2, see the AWS SDK for Java, version 2 guide.

Breaking changes

Note the following breaking changes:

- In AWS Glue 5.0, when using the S3A file system and if both `fs.s3a.endpoint` and `fs.s3a.endpoint.region` are not set, the default region used by S3A is `us-east-2`. This can cause issues, such as S3 upload timeout errors, especially for VPC jobs. To mitigate the issues caused by this change, set the `` `fs.s3a.endpoint.region` `` Spark configuration when using the S3A file system in AWS Glue 5.0.

- Lake Formation Fine-grained Access Control (FGAC)
  - AWS Glue 5.0 only supports the new Spark-native FGAC using Spark DataFrames. It does not support FGAC using AWS Glue DynamicFrames.
    - Use of FGAC in 5.0 requires migrating from AWS Glue DynamicFrames to Spark DataFrames
    - If you don't need FGAC, then it is not necessary to migrate to Spark DataFrame and GlueContext features, like job bookmarks and push down predicates, will continue to work.
  - Jobs with Spark-native FGAC require a minimum of 4 workers: one user driver, one system driver, one system executor, and one standby user executor.
  - For more information, see Using AWS Glue with AWS Lake Formation for fine-grained access control.
- Lake Formation Full Table Access (FTA)
  - AWS Glue 5.0 supports FTA with Spark-native DataFrames (new) and GlueContext DynamicFrames (legacy, with limitations)
  - Spark-native FTA
    - If 4.0 script uses GlueContext, migrate to using native spark.
    - This feature is limited to hive and iceberg tables
    - For more info on configuring a 5.0 job to use Spark native FTA, see
  - GlueContext DynamicFrame FTA
    - No code change necessary
    - This feature is limited to non-OTF tables - it will not work with Iceberg, Delta Lake, and Hudi.
- Vectorized SIMD CSV reader is not supported.
- Continuous logging to output log group is not supported. Use `error` log group instead.
- The AWS Glue job run insights `job-insights-rule-driver` has been deprecated. The `job-insights-rca-driver` log stream is now located in the error log group.
- Athena-based custom/marketplace connectors are not supported.
- Adobe Marketo Engage, Facebook Ads, Google Ads, Google Analytics 4, Google Sheets, Hubspot, Instagram Ads, Intercom, Jira Cloud, Oracle NetSuite, Salesforce, Salesforce Marketing Cloud, Salesforce Marketing Cloud Account Engagement, SAP OData, ServiceNow, Slack, Snapchat Ads, Stripe, Zendesk and Zoho CRM connectors are not supported.

- Custom log4j properties are not supported in AWS Glue 5.0.

Major enhancements from Spark 3.3.0 to Spark 3.5.4

Note the following enhancements:

- Python client for Spark Connect (SPARK-39375).
- Implement support for DEFAULT values for columns in tables (SPARK-38334).
- Support "Lateral Column Alias References" (SPARK-27561).
- Harden SQLSTATE usage for error classes (SPARK-41994).
- Enable Bloom filter Joins by default (SPARK-38841).
- Better Spark UI scalability and driver stability for large applications (SPARK-41053).
- Async Progress Tracking in Structured Streaming (SPARK-39591).
- Python arbitrary stateful processing in structured streaming (SPARK-40434).
- Pandas API coverage improvements (SPARK-42882) and NumPy input support in PySpark (SPARK-39405).
- Provide a memory profiler for PySpark user-defined functions (SPARK-40281).
- Implement PyTorch distributor (SPARK-41589).
- Publish SBOM artifacts (SPARK-41893).
- Support IPv6-only environment (SPARK-39457).
- Customized K8s scheduler (Apache YuniKorn and Volcano) GA (SPARK-42802).
- Scala and Go client support in Spark Connect (SPARK-42554) and (SPARK-43351).
- PyTorch-based distributed ML Support for Spark Connect (SPARK-42471).
- Structured streaming support for Spark Connect in Python and Scala (SPARK-42938).
- Pandas API support for the Python Spark Connect Client (SPARK-42497).
- Introduce Arrow Python UDFs (SPARK-40307).
- Support Python user-defined table functions (SPARK-43798).
- Migrate PySpark errors onto error classes (SPARK-42986).
- PySpark test framework (SPARK-44042).
- Add support for Datasketches HllSketch (SPARK-16484).
- Built-in SQL function improvement (SPARK-41231).
- IDENTIFIER clause (SPARK-43205).

- Add SQL functions into Scala, Python and R API (SPARK-43907).
- Add named argument support for SQL functions (SPARK-43922).
- Avoid unnecessary task rerun on decommissioned executor lost if shuffle data migrated (SPARK-41469).
- Distributed ML <> spark connect (SPARK-42471).
- DeepSpeed distributor (SPARK-44264).
- Implement changelog checkpointing for RocksDB state store (SPARK-43421).
- Introduce watermark propagation among operators (SPARK-42376).
- Introduce dropDuplicatesWithinWatermark (SPARK-42931).
- RocksDB state store provider memory management enhancements (SPARK-43311).

Actions to migrate to AWS Glue 5.0

For existing jobs, change the `Glue version` from the previous version to `Glue 5.0` in the job configuration.

- In AWS Glue Studio, choose `Glue 5.0 - Supports Spark 3.5.4, Scala 2, Python 3` in `Glue version`.
- In the API, choose **`5.0`** in the `GlueVersion` parameter in the `UpdateJob` API operation.

For new jobs, choose `Glue 5.0` when you create a job.

- In the console, choose `Spark 3.5.4, Python 3 (Glue Version 5.0) or Spark 3.5.4, Scala 2 (Glue Version 5.0)` in `Glue version`.
- In AWS Glue Studio, choose `Glue 5.0 - Supports Spark 3.5.4, Scala 2, Python 3` in `Glue version`.
- In the API, choose **`5.0`** in the `GlueVersion` parameter in the `CreateJob` API operation.

To view Spark event logs of AWS Glue 5.0 coming from AWS Glue 2.0 or earlier, launch an upgraded Spark history server for AWS Glue 5.0 using AWS CloudFormation or Docker.

Migration checklist

Review this checklist for migration:

- Java 17 updates
- [Scala] Upgrade AWS SDK calls from v1 to v2
- Python 3.10 to 3.11 migration
- [Python] Update boto references from 1.26 to 1.34

AWS Glue 5.0 features

This section describes AWS Glue features in more detail.

Querying metastore data catalogs from AWS Glue ETL

You can register your AWS Glue job to access the AWS Glue Data Catalog, which makes tables and other metastore resources available to disparate consumers. The Data Catalog supports a multi-catalog hierarchy, which unifies all your data across Amazon S3 data lakes. It also provides both a Hive metastore API and an open-source Apache Iceberg API for accessing the data. These features are available to AWS Glue and other data-oriented services like Amazon EMR, Amazon Athena and Amazon Redshift.

When you create resources in the Data Catalog, you can access them from any SQL engine that supports the Apache Iceberg REST API. AWS Lake Formation manages permissions. Following configuration, you can leverage AWS Glue's capabilities to query disparate data by querying these metastore resources with familiar applications. These include Apache Spark and Trino.

How metadata resources are organized

Data is organized in a logical hierarchy of catalogs, databases and tables, using the AWS Glue Data Catalog:

- Catalog – A logical container that holds objects from a data store, such as schemas or tables.
- Database – Organizes data objects such as tables and views in a catalog.
- Tables and views – Data objects in a database that provide an abstraction layer with an understandable schema. They make it easy to access underlying data, which could be in various formats and in various locations.

Migrating from AWS Glue 4.0 to AWS Glue 5.0

All existing job parameters and major features that exist in AWS Glue 4.0 will exist in AWS Glue 5.0, except machine learning transforms.

The following new parameters were added:

- `--enable-lakeformation-fine-grained-access`: Enables the fine-grained access control (FGAC) feature in AWS Lake Formation tables.

Refer to the Spark migration documentation:

- Migration Guide: Spark Core
- Migration Guide: SQL, Datasets and DataFrame
- Migration Guide: Structured Streaming
- Upgrading PySpark

Migrating from AWS Glue 3.0 to AWS Glue 5.0

> **Note**
>
> For migration steps related to AWS Glue 4.0, see Migrating from AWS Glue 3.0 to AWS Glue 4.0.

All existing job parameters and major features that exist in AWS Glue 3.0 will exist in AWS Glue 5.0, except machine learning transforms.

Migrating from AWS Glue 2.0 to AWS Glue 5.0

> **Note**
>
> For migration steps related to AWS Glue 4.0 and a list of migration differences between AWS Glue version 3.0 and 4.0, see Migrating from AWS Glue 3.0 to AWS Glue 4.0.

Also note the following migration differences between AWS Glue versions 3.0 and 2.0:

- All existing job parameters and major features that exist in AWS Glue 2.0 will exist in AWS Glue 5.0, except machine learning transforms.

- Several Spark changes alone may require revision of your scripts to ensure removed features are not being referenced. For example, Spark 3.1.1 and later does not enable Scala-untyped UDFs but Spark 2.4 does allow them.
- Python 2.7 is not supported.
- Any extra jars supplied in existing AWS Glue 2.0 jobs may bring in conflicting dependencies since there were upgrades in several dependencies. You can avoid classpath conflicts with the `--user-jars-first` job parameter.
- Changes to the behavior of loading/saving of timestamps from/to parquet files. For more details, see Upgrading from Spark SQL 3.0 to 3.1.
- Different Spark task parallelism for driver/executor configuration. You can adjust task parallelism by passing the `--executor-cores` job argument.

Logging behavior changes in AWS Glue 5.0

The following are changes in logging behavior in AWS Glue 5.0. For more information, see Logging for AWS Glue jobs.

- All logs (system logs, Spark daemon logs, user logs, and Glue Logger logs) are now written to the `/aws-glue/jobs/error` log group by default.
- The `/aws-glue/jobs/logs-v2` log group used for continuous logging in previous versions is no longer used.
- You can no longer rename or customize the log group or log stream names using the removed continuous logging arguments. Instead, see the new job arguments in AWS Glue 5.0.

Two new job arguments are introduced in AWS Glue 5.0

- `--custom-logGroup-prefix`: Allows you to specify a custom prefix for the `/aws-glue/jobs/error` and `/aws-glue/jobs/output` log groups.
- `--custom-logStream-prefix`: Allows you to specify a custom prefix for the log stream names within the log groups.

  Validation rules and limitations for custom prefixes include:

  - The entire log stream name must be between 1 and 512 characters.
  - The custom prefix for log stream names is limited to 400 characters.

- Allowed characters in prefixes include alphanumeric characters, underscores (`_`), hyphens (`-`), and forward slashes (`/`).

Deprecated continuous logging arguments in AWS Glue 5.0

The following job arguments for continuous logging have been deprecated in AWS Glue 5.0

- `--enable-continuous-cloudwatch-log`
- `--continuous-log-logGroup`
- `--continuous-log-logStreamPrefix`
- `--continuous-log-conversionPattern`
- `--enable-continuous-log-filter`

Connector and JDBC driver migration for AWS Glue 5.0

For the versions of JDBC and data lake connectors that were upgraded, see:

- Appendix B: JDBC driver upgrades
- Appendix C: Connector upgrades
- Appendix D: Open table format upgrades

The following changes apply to the connector or driver versions identified in the appendices for Glue 5.0.

Amazon Redshift

Note the following changes:

- Adds support for three-part table names to allow connector to query Redshift data sharing tables.
- Corrects mapping of Spark `ShortType` to use Redshift `SMALLINT` instead of `INTEGER` to better match expected data size.
- Added support for Custom Cluster Names (CNAME) for Amazon Redshift Serverless.

Apache Hudi

Note the following changes:

- Support record level index.
- Support auto generation of record keys. Now you don't have to specify the record key field.

Apache Iceberg

Note the following changes:

- Support fine-grained access control with AWS Lake Formation.
- Support branching and tagging which are named references to snapshots with their own independent lifecycles.
- Added a changelog view procedure which generates a view that contains the changes made to a table over a specified period or between specific snapshots.

Delta Lake

Note the following changes:

- Support Delta Universal Format (UniForm) which enables seamless access through Apache Iceberg and Apache Hudi.
- Support Deletion Vectors that implements a Merge-on-Read paradigm.

AzureCosmos

Note the following changes:

- Added hierarchical partition key support.
- Added option to use custom Schema with StringType (raw json) for a nested property.
- Added config option `spark.cosmos.auth.aad.clientCertPemBase64` to allow using SPN (ServicePrincipal name) authentication with certificate instead of client secret.

For more information, see the Azure Cosmos DB Spark connector change log.

Microsoft SQL Server

Note the following changes:

- TLS encryption is enabled by default.
- When encrypt = false but the server requires encryption, the certificate is validated based on the `trustServerCertificate` connection setting.
- `aadSecurePrincipalId` and `aadSecurePrincipalSecret` deprecated.
- `getAADSecretPrincipalId` API removed.
- Added CNAME resolution when realm is specified.

MongoDB

Note the following changes:

- Support for micro-batch mode with Spark Structured Streaming.
- Support for BSON data types.
- Added support for reading multiple collections when using micro-batch or continuous streaming modes.
  - If the name of a collection used in your `collection` configuration option contains a comma, the Spark Connector treats it as two different collections. To avoid this, you must escape the comma by preceding it with a backslash (\).
  - If the name of a collection used in your `collection` configuration option is "*", the Spark Connector interprets it as a specification to scan all collections. To avoid this, you must escape the asterisk by preceding it with a backslash (\).
  - If the name of a collection used in your `collection` configuration option contains a backslash (\), the Spark Connector treats the backslash as an escape character, which might change how it interprets the value. To avoid this, you must escape the backslash by preceding it with another backslash.

For more information, see the MongoDB connector for Spark release notes.

Snowflake

Note the following changes:

- Introduced a new `trim_space` parameter that you can use to trim values of `StringType` columns automatically when saving to a Snowflake table. Default: `false`.
- Disabled the `abort_detached_query` parameter at the session level by default.
- Removed the requirement of the `SFUSER` parameter when using OAUTH.

- Removed the Advanced Query Pushdown feature. Alternatives to the feature are available. For example, instead of loading data from Snowflake tables, users can directly load data from Snowflake SQL queries.

For more information, see the Snowflake Connector for Spark release notes.

Appendix A: Notable dependency upgrades

The following are dependency upgrades:

Dependency	Version in AWS Glue 5.0	Version in AWS Glue 4.0	Version in AWS Glue 3.0	Version in AWS Glue 2.0	Version in AWS Glue 1.0
Java	17	8	8	8	8
Spark	3.5.4	3.3.0-amzn-1	3.1.1-amzn-0	2.4.3	2.4.3
Hadoop	3.4.1	3.3.3-amzn-0	3.2.1-amzn-3	2.8.5-amzn-5	2.8.5-amzn-1
Scala	2.12.18	2.12	2.12	2.11	2.11
Jackson	2.15.2	2.12	2.12	2.11	2.11
Hive	2.3.9-amzn-4	2.3.9-amzn-2	2.3.7-amzn-4	1.2	1.2
EMRFS	2.69.0	2.54.0	2.46.0	2.38.0	2.30.0
Json4s	3.7.0-M11	3.7.0-M11	3.6.6	3.5.x	3.5.x
Arrow	12.0.1	7.0.0	2.0.0	0.10.0	0.10.0
AWS Glue Data Catalog client	4.5.0	3.7.0	3.0.0	1.10.0	N/A
AWS SDK for Java	2.29.52	1.12	1.12		
Python	3.11	3.10	3.7	2.7 & 3.6	2.7 & 3.6

Dependency	Version in AWS Glue 5.0	Version in AWS Glue 4.0	Version in AWS Glue 3.0	Version in AWS Glue 2.0	Version in AWS Glue 1.0
Boto	1.34.131	1.26	1.18	1.12	N/A
EMR DynamoDB connector	5.6.0	4.16.0			

Appendix B: JDBC driver upgrades

The following are JDBC driver upgrades:

Driver	JDBC driver version in AWS Glue 5.0	JDBC driver version in AWS Glue 4.0	JDBC driver version in AWS Glue 3.0	JDBC driver version in past AWS Glue versions
MySQL	8.0.33	8.0.23	8.0.23	5.1
Microsoft SQL Server	10.2.0	9.4.0	7.0.0	6.1.0
Oracle Databases	23.3.0.23.09	21.7	21.1	11.2
PostgreSQL	42.7.3	42.3.6	42.2.18	42.1.0
Amazon Redshift	redshift-jdbc42-2.1.0.29	redshift-jdbc42-2.1.0.16	redshift-jdbc41-1.2.12.1017	redshift-jdbc41-1.2.12.1017
SAP Hana	2.20.17	2.17.12		
Teradata	20.00.00.33	20.00.00.06		

Appendix C: Connector upgrades

The following are connector upgrades:

Driver	Connector version in AWS Glue 5.0	Connector version in AWS Glue 4.0	Connector version in AWS Glue 3.0
EMR DynamoDB connector	5.6.0	4.16.0	
Amazon Redshift	6.4.0	6.1.3	
OpenSearch	1.2.0	1.0.1	
MongoDB	10.4.0	10.0.4	3.0.0
Snowflake	3.0.0	2.12.0	
Google BigQuery	0.32.2	0.32.2	
AzureCosmos	4.33.0	4.22.0	
AzureSQL	1.3.0	1.3.0	
Vertica	3.3.5	3.3.5	

Appendix D: Open table format upgrades

The following are open table format upgrades:

OTF	Connector version in AWS Glue 5.0	Connector version in AWS Glue 4.0	Connector version in AWS Glue 3.0
Hudi	0.15.0	0.12.1	0.10.1
Delta Lake	3.3.0	2.1.0	1.0.0
Iceberg	1.7.1	1.0.0	0.13.1

Migrating AWS Glue for Spark jobs to AWS Glue version 4.0

This topic describes the changes between AWS Glue versions 0.9, 1.0, 2.0, and 3.0 to allow you to migrate your Spark applications and ETL jobs to AWS Glue 4.0. It also describes the features in AWS Glue 4.0 and the advantages of using it.

To use this feature with your AWS Glue ETL jobs, choose **`4.0`** for the `Glue version` when creating your jobs.

Topics

- New features supported
- Actions to migrate to AWS Glue 4.0
- Migration checklist
- Migrating from AWS Glue 3.0 to AWS Glue 4.0
- Migrating from AWS Glue 2.0 to AWS Glue 4.0
- Migrating from AWS Glue 1.0 to AWS Glue 4.0
- Migrating from AWS Glue 0.9 to AWS Glue 4.0
- Connector and JDBC driver migration for AWS Glue 4.0
- Appendix A: Notable dependency upgrades
- Appendix B: JDBC driver upgrades
- Appendix C: Connector upgrades

New features supported

This section describes new features and advantages of AWS Glue version 4.0.

- It is based on Apache Spark 3.3.0, but includes optimizations in AWS Glue, and Amazon EMR, such as adaptive query runs, vectorized readers, and optimized shuffles and partition coalescing.
- Upgraded JDBC drivers for all AWS Glue native sources including MySQL, Microsoft SQL Server, Oracle, PostgreSQL, MongoDB, and upgraded Spark libraries and dependencies brought in by Spark 3.3.0.
- Updated with a new Amazon Redshift connector and JDBC driver.
- Optimized Amazon S3 access with upgraded EMR File System (EMRFS) and enabled Amazon S3-optimized output committers, by default.

- Optimized Data Catalog access with partition indexes, pushdown predicates, partition listing, and an upgraded Hive metastore client.
- Integration with Lake Formation for governed catalog tables with cell-level filtering and data lake transactions.
- Reduced startup latency to improve overall job completion times and interactivity.
- Spark jobs are billed in 1-second increments with a 10x lower minimum billing duration—from a 10-minute minimum to a 1-minute minimum.
- Native support for open-data lake frameworks with Apache Hudi, Delta Lake, and Apache Iceberg.
- Native support for the Amazon S3-based Cloud Shuffle Storage Plugin (an Apache Spark plugin) to use Amazon S3 for shuffling and elastic storage capacity.

Major enhancements from Spark 3.1.1 to Spark 3.3.0

Note the following enhancements:

- Row-level runtime filtering (SPARK-32268).
- ANSI enhancements (SPARK-38860).
- Error message improvements (SPARK-38781).
- Support complex types for Parquet vectorized reader (SPARK-34863).
- Hidden file metadata support for Spark SQL (SPARK-37273).
- Provide a profiler for Python/Pandas UDFs (SPARK-37443).
- Introduce Trigger.AvailableNow for running streaming queries like Trigger.Once in multiple batches (SPARK-36533).
- More comprehensive Datasource V2 pushdown capabilities (SPARK-38788).
- Migrating from log4j 1 to log4j 2 (SPARK-37814).

Other notable changes

Note the following changes:

- Breaking changes
  - Drop references to Python 3.6 support in docs and Python/docs (SPARK-36977).
  - Remove named tuple hack by replacing built-in pickle to cloudpickle (SPARK-32079).

  - Bump minimum pandas version to 1.0.5 (SPARK-37465).

Actions to migrate to AWS Glue 4.0

For existing jobs, change the `Glue version` from the previous version to `Glue 4.0` in the job configuration.

- In AWS Glue Studio, choose `Glue 4.0 - Supports Spark 3.3, Scala 2, Python 3` in `Glue version`.
- In the API, choose **`4.0`** in the `GlueVersion` parameter in the `UpdateJob` API operation.

For new jobs, choose `Glue 4.0` when you create a job.

- In the console, choose `Spark 3.3, Python 3 (Glue Version 4.0) or Spark 3.3, Scala 2 (Glue Version 3.0)` in `Glue version`.
- In AWS Glue Studio, choose `Glue 4.0 - Supports Spark 3.3, Scala 2, Python 3` in `Glue version`.
- In the API, choose **`4.0`** in the `GlueVersion` parameter in the `CreateJob` API operation.

To view Spark event logs of AWS Glue 4.0 coming from AWS Glue 2.0 or earlier, launch an upgraded Spark history server for AWS Glue 4.0 using AWS CloudFormation or Docker.

Migration checklist

- Do your job's external Python libraries depend on Python 2.7/3.6?
  - Update the dependent libraries from Python 2.7/3.6 to Python 3.10 as Spark 3.3.0 completely removed Python 2.7 and 3.6 support.

Migrating from AWS Glue 3.0 to AWS Glue 4.0

Note the following changes when migrating:

- All existing job parameters and major features that exist in AWS Glue 3.0 will exist in AWS Glue 4.0.
- AWS Glue 3.0 uses Amazon EMR-optimized Spark 3.1.1, and AWS Glue 4.0 uses Amazon EMR-optimized Spark 3.3.0.

Several Spark changes alone might require revision of your scripts to ensure that removed features are not being referenced.

- AWS Glue 4.0 also features an update to EMRFS and Hadoop. For the specific version, see Appendix A: Notable dependency upgrades.
- The AWS SDK provided in ETL jobs is now upgraded from 1.11 to 1.12.
- All Python jobs will be using Python version 3.10. Previously, Python 3.7 was used in AWS Glue 3.0.

  As a result, some pymodules brought out-of-the-box by AWS Glue are upgraded.
- Log4j has been upgraded to Log4j2.
  - For information on the Log4j2 migration path, see the Log4j documentation.
  - You must rename any custom log4j.properties file as a log4j2.properties file instead, with the appropriate log4j2 properties.
- For migrating certain connectors, see Connector and JDBC driver migration for AWS Glue 4.0.
- The AWS Encryption SDK is upgraded from 1.x to 2.x. AWS Glue jobs using AWS Glue security configurations and jobs dependent on the AWS Encryption SDK dependency provided in runtime are affected. See the instructions for AWS Glue job migration.

  You can safely upgrade an AWS Glue 2.0/3.0 job to an AWS Glue 4.0 job because AWS Glue 2.0/3.0 already contains the AWS Encryption SDK bridge version.

Refer to the Spark migration documentation:

- Upgrading from Spark SQL 3.1 to 3.2
- Upgrading from Spark SQL 3.2 to 3.3

Migrating from AWS Glue 2.0 to AWS Glue 4.0

Note the following changes when migrating:

> **Note**
>
> For migration steps related to AWS Glue 3.0, see Migrating from AWS Glue 3.0 to AWS Glue 4.0.

- All existing job parameters and major features that exist in AWS Glue 2.0 will exist in AWS Glue 4.0.
- The EMRFS S3-optimized committer for writing Parquet data into Amazon S3 is enabled by default since AWS Glue 3.0. However, you can still disable it by setting `--enable-s3-parquet-optimized-committer` to `false`.
- AWS Glue 2.0 uses open-source Spark 2.4 and AWS Glue 4.0 uses Amazon EMR-optimized Spark 3.3.0.
  - Several Spark changes alone may require revision of your scripts to ensure that removed features are not being referenced.
  - For example, Spark 3.3.0 does not enable Scala-untyped UDFs, but Spark 2.4 does allow them.
- The AWS SDK provided in ETL jobs is now upgraded from 1.11 to 1.12.
- AWS Glue 4.0 also features an update to EMRFS, updated JDBC drivers, and inclusions of additional optimizations onto Spark itself provided by AWS Glue.
- Scala is updated to 2.12 from 2.11, and Scala 2.12 is not backward compatible with Scala 2.11.
- Python 3.10 is the default version used for Python scripts, as AWS Glue 2.0 was only using Python 3.7 and 2.7.
  - Python 2.7 is not supported with Spark 3.3.0. Any job requesting Python 2 in the job configuration will fail with an IllegalArgumentException.
  - A new mechanism of installing additional Python modules is available since AWS Glue 2.0.
- Several dependency updates, highlighted in Appendix A: Notable dependency upgrades.
- Any extra JAR files supplied in existing AWS Glue 2.0 jobs might bring in conflicting dependencies because there were upgrades in several dependencies in 4.0 from 2.0. You can avoid classpath conflicts in AWS Glue 4.0 with the `--user-jars-first` AWS Glue job parameter.
- AWS Glue 4.0 uses Spark 3.3. Starting with Spark 3.1, there was a change in the behavior of loading/saving of timestamps from/to parquet files. For more details, see Upgrading from Spark SQL 3.0 to 3.1.

  We recommend to set the following parameters when reading/writing parquet data that contains timestamp columns. Setting those parameters can resolve the calendar incompatibility issue that occurs during the Spark 2 to Spark 3 upgrade, for both the AWS Glue Dynamic Frame and Spark Data Frame. Use the CORRECTED option to read the datetime value as it is; and the LEGACY option to rebase the datetime values with regard to the calendar difference during reading.

```
- Key: --conf
- Value: spark.sql.legacy.parquet.int96RebaseModeInRead=[CORRECTED|LEGACY] --
conf spark.sql.legacy.parquet.int96RebaseModeInWrite=[CORRECTED|LEGACY] --conf
 spark.sql.legacy.parquet.datetimeRebaseModeInRead=[CORRECTED|LEGACY]
```

- For migrating certain connectors, see Connector and JDBC driver migration for AWS Glue 4.0.
- The AWS Encryption SDK is upgraded from 1.x to 2.x. AWS Glue jobs using AWS Glue security configurations and jobs dependent on the AWS Encryption SDK dependency provided in runtime are affected. See these instructions for AWS Glue job migration:
    - You can safely upgrade an AWS Glue 2.0 job to an AWS Glue 4.0 job because AWS Glue 2.0 already contains the AWS Encryption SDK bridge version.

Refer to the Spark migration documentation:

- Upgrading from Spark SQL 2.4 to 3.0
- Upgrading from Spark SQL 3.1 to 3.2
- Upgrading from Spark SQL 3.2 to 3.3
- Changes in Datetime behavior to be expected since Spark 3.0.

Migrating from AWS Glue 1.0 to AWS Glue 4.0

Note the following changes when migrating:

- AWS Glue 1.0 uses open-source Spark 2.4 and AWS Glue 4.0 uses Amazon EMR-optimized Spark 3.3.0.
    - Several Spark changes alone may require revision of your scripts to ensure that removed features are not being referenced.
    - For example, Spark 3.3.0 does not enable Scala-untyped UDFs, but Spark 2.4 does allow them.
- All jobs in AWS Glue 4.0 will be run with significantly improved startup times. Spark jobs will be billed in 1-second increments with a 10x lower minimum billing duration since startup latency will go from 10 minutes maximum to 1 minute maximum.
- Logging behavior has changed significantly in AWS Glue 4.0, Spark 3.3.0 has a minimum requirement of Log4j2.
- Several dependency updates, highlighted in the appendix.

- Scala is also updated to 2.12 from 2.11, and Scala 2.12 is not backward compatible with Scala 2.11.
- Python 3.10 is also the default version used for Python scripts, as AWS Glue 0.9 was only using Python 2.

  Python 2.7 is not supported with Spark 3.3.0. Any job requesting Python 2 in the job configuration will fail with an IllegalArgumentException.
- A new mechanism of installing additional Python modules through pip is available since AWS Glue 2.0. For more information, see Installing additional Python modules with pip in AWS Glue 2.0+.
- AWS Glue 4.0 does not run on Apache YARN, so YARN settings do not apply.
- AWS Glue 4.0 does not have a Hadoop Distributed File System (HDFS).
- Any extra JAR files supplied in existing AWS Glue 1.0 jobs might bring in conflicting dependencies because there were upgrades in several dependencies in 4.0 from 1.0. We enable AWS Glue 4.0 with the `--user-jars-first` AWS Glue job parameter by default, to avoid this problem.
- AWS Glue 4.0 supports auto scaling. Therefore, the ExecutorAllocationManager metric will be available when auto scaling is enabled.
- In AWS Glue version 4.0 jobs, you specify the number of workers and worker type, but do not specify a `maxCapacity`.
- AWS Glue 4.0 does not yet support machine learning transforms.
- For migrating certain connectors, see Connector and JDBC driver migration for AWS Glue 4.0.
- The AWS Encryption SDK is upgraded from 1.x to 2.x. AWS Glue jobs using AWS Glue security configurations and jobs dependent on the AWS Encryption SDK dependency provided in runtime are affected. See these instructions for AWS Glue job migration.
  - You cannot migrate an AWS Glue 0.9/1.0 job to an AWS Glue 4.0 job directly. This is because when upgrading directly to version 2.x or later and enabling all new features immediately, the AWS Encryption SDK won't be able to decrypt the ciphertext encrypted under earlier versions of the AWS Encryption SDK.
  - To safely upgrade, we first recommend that you migrate to an AWS Glue 2.0/3.0 job that contains the AWS Encryption SDK bridge version. Run the job once to utilize the AWS Encryption SDK bridge version.
  - Upon completion, you can safely migrate the AWS Glue 2.0/3.0 job to AWS Glue 4.0.

Refer to the Spark migration documentation:

- Upgrading from Spark SQL 2.4 to 3.0
- Upgrading from Spark SQL 3.0 to 3.1
- Upgrading from Spark SQL 3.1 to 3.2
- Upgrading from Spark SQL 3.2 to 3.3
- Changes in Datetime behavior to be expected since Spark 3.0.

Migrating from AWS Glue 0.9 to AWS Glue 4.0

Note the following changes when migrating:

- AWS Glue 0.9 uses open-source Spark 2.2.1 and AWS Glue 4.0 uses Amazon EMR-optimized Spark 3.3.0.
  - Several Spark changes alone might require revision of your scripts to ensure that removed features are not being referenced.
  - For example, Spark 3.3.0 does not enable Scala-untyped UDFs, but Spark 2.2 does allow them.
- All jobs in AWS Glue 4.0 will be run with significantly improved startup times. Spark jobs will be billed in 1-second increments with a 10x lower minimum billing duration because startup latency will go from 10 minutes maximum to 1 minute maximum.
- Logging behavior has changed significantly since AWS Glue 4.0, Spark 3.3.0 has a minimum requirement of Log4j2 as mentioned here (https://spark.apache.org/docs/latest/core-migration-guide.html#upgrading-from-core-32-to-33).
- Several dependency updates, highlighted in the appendix.
- Scala is also updated to 2.12 from 2.11, and Scala 2.12 is not backward compatible with Scala 2.11.
- Python 3.10 is also the default version used for Python scripts, as AWS Glue 0.9 was only using Python 2.
  - Python 2.7 is not supported with Spark 3.3.0. Any job requesting Python 2 in the job configuration will fail with an IllegalArgumentException.
  - A new mechanism of installing additional Python modules through pip is available.
- AWS Glue 4.0 does not run on Apache YARN, so YARN settings do not apply.
- AWS Glue 4.0 does not have a Hadoop Distributed File System (HDFS).

- Any extra JAR files supplied in existing AWS Glue 0.9 jobs might bring in conflicting dependencies because there were upgrades in several dependencies in 3.0 from 0.9. You can avoid classpath conflicts in AWS Glue 3.0 with the `--user-jars-first` AWS Glue job parameter.
- AWS Glue 4.0 supports auto scaling. Therefore, the ExecutorAllocationManager metric will be available when auto scaling is enabled.
- In AWS Glue version 4.0 jobs, you specify the number of workers and worker type, but do not specify a `maxCapacity`.
- AWS Glue 4.0 does not yet support machine learning transforms.
- For migrating certain connectors, see Connector and JDBC driver migration for AWS Glue 4.0.
- The AWS Encryption SDK is upgraded from 1.x to 2.x. AWS Glue jobs using AWS Glue security configurations and jobs dependent on the AWS Encryption SDK dependency provided in runtime are affected. See these instructions for AWS Glue job migration.
  - You cannot migrate an AWS Glue 0.9/1.0 job to an AWS Glue 4.0 job directly. This is because when upgrading directly to version 2.x or later and enabling all new features immediately, the AWS Encryption SDK won't be able to decrypt the ciphertext encrypted under earlier versions of the AWS Encryption SDK.
  - To safely upgrade, we first recommend that you migrate to an AWS Glue 2.0/3.0 job that contains the AWS Encryption SDK bridge version. Run the job once to utilize the AWS Encryption SDK bridge version.
  - Upon completion, you can safely migrate the AWS Glue 2.0/3.0 job to AWS Glue 4.0.

Refer to the Spark migration documentation:

- Upgrading from Spark SQL 2.2 to 2.3
- Upgrading from Spark SQL 2.3 to 2.4
- Upgrading from Spark SQL 2.4 to 3.0
- Upgrading from Spark SQL 3.0 to 3.1
- Upgrading from Spark SQL 3.1 to 3.2
- Upgrading from Spark SQL 3.2 to 3.3
- Changes in Datetime behavior to be expected since Spark 3.0.

Connector and JDBC driver migration for AWS Glue 4.0

For the versions of JDBC and data lake connectors that were upgraded, see:

- Appendix B: JDBC driver upgrades
- Appendix C: Connector upgrades

Hudi

- Spark SQL support improvements:
  - Through the `Call Procedure` command, there is added support for upgrade, downgrade, bootstrap, clean, and repair. `Create/Drop/Show/Refresh Index` syntax is possible in Spark SQL.
  - A performance gap has been closed between usage through a Spark DataSource as opposed to Spark SQL. Datasource writes in the past used to be faster than SQL.
  - All built-in key generators implement more performant Spark-specific API operations.
  - Replaced UDF transformation in the bulk `insert` operation with RDD transformations to cut down on costs of using SerDe.
  - Spark SQL with Hudi requires a `primaryKey` to be specified by `tblproperites` or options in the SQL statement. For update and delete operations, the `preCombineField` is required as well.
- Any Hudi table created before version 0.10.0 without a `primaryKey` needs to be recreated with a `primaryKey` field since version 0.10.0.

PostgreSQL

- Several vulnerabilities (CVEs) were addressed.
- Java 8 is natively supported.
- If the job is using Arrays of Arrays, with the exception of byte arrays, this scenario can be treated as multidimensional arrays.

MongoDB

- The current MongoDB connector supports Spark version 3.1 or later and MongoDB version 4.0 or later.

- Due to the connector upgrade, a few property names changed. For example, the URI property name changed to `connection.uri`. For more information on the current options, see the MongoDB Spark Connector blog.
- Using MongoDB 4.0 hosted by Amazon DocumentDB has some functional differences. For more information, see these topics:
  - Functional Differences: Amazon DocumentDB and MongoDB
  - Supported MongoDB APIs, Operations, and Data Types.
- The "partitioner" option is restricted to `ShardedPartitioner`, `PaginateIntoPartitionsPartitioner`, and `SinglePartitionPartitioner`. It cannot use default `SamplePartitioner` and `PaginateBySizePartitioner` for Amazon DocumentDB because the stage operator does not support the MongoDB API. For more information, see Supported MongoDB APIs, Operations, and Data Types.

Delta Lake

- Delta Lake now supports time travel in SQL to query older data easily. With this update, time travel is now available both in Spark SQL and through the DataFrame API. Support has been added for the current version of TIMESTAMP in SQL.
- Spark 3.3 introduces Trigger.AvailableNow for running streaming queries as an equivalent to `Trigger.Once` for batch queries. This support is also available when using Delta tables as a streaming source.
- Support for SHOW COLUMNS to return the list of columns in a table.
- Support for DESCRIBE DETAIL in the Scala and Python DeltaTable API. It retrieves detailed information about a Delta table using either the DeltaTable API or Spark SQL.
- Support for returning operation metrics from SQL Delete, Merge, and Update commands. Previously these SQL commands returned an empty DataFrame, now they return a DataFrame with useful metrics about the operation performed.
- Optimize performance improvements:
  - Set the configuration option `spark.databricks.delta.optimize.repartition.enabled=true` to use `repartition(1)` instead of `coalesce(1)` in the Optimize command for better performance when compacting many small files.
  - Improved performance by using a queue-based approach to parallelize compaction jobs.
- Other notable changes:

- Support for using variables in the VACUUM and OPTIMIZE SQL commands.
- Improvements for CONVERT TO DELTA with catalog tables including:
    - Autofill the partition schema from the catalog when it's not provided.
    - Use partition information from the catalog to find the data files to commit instead of doing a full directory scan. Instead of committing all data files in the table directory, only data files under the directories of active partitions will be committed.
- Support for Change Data Feed (CDF) batch reads on column mapping enabled tables when DROP COLUMN and RENAME COLUMN have not been used. For more information, see the Delta Lake documentation.
- Improve Update command performance by enabling schema pruning in the first pass.

Apache Iceberg

- Added several performance improvements for scan planning and Spark queries.
- Added a common REST catalog client that uses change-based commits to resolve commit conflicts on the service side.
- `AS OF` syntax for SQL time travel queries is supported.
- Added merge-on-read support for MERGE and UPDATE queries.
- Added support to rewrite partitions using Z-order.
- Added a spec and implementation for Puffin, a format for large stats and index blobs, like Theta sketches or bloom filters.
- Added new interfaces for consuming data incrementally (both append and changelog scans).
- Added support for bulk operations and ranged reads to FileIO interfaces.
- Added more metadata tables to show delete files in the metadata tree.
- The drop table behavior changed. In Iceberg 0.13.1, running `DROP TABLE` removes the table from the catalog and deletes the table contents as well. In Iceberg 1.0.0, `DROP TABLE` only removes the table from the catalog. To delete the table contents use `DROP TABLE PURGE`.
- Parquet vectorized reads are enabled by default in Iceberg 1.0.0. If you want to disable vectorized reads, set `read.parquet.vectorization.enabled` to `false`.

Oracle

Changes are minor.

MySQL

Changes are minor.

Amazon Redshift

AWS Glue 4.0 features a new Amazon Redshift connector with a new JDBC driver. For information about the enhancements and how to migrate from previous AWS Glue versions, see the section called “Redshift connections”.

Appendix A: Notable dependency upgrades

The following are dependency upgrades:

Dependency	Version in AWS Glue 4.0	Version in AWS Glue 3.0	Version in AWS Glue 2.0	Version in AWS Glue 1.0
Spark	3.3.0-amzn-1	3.1.1-amzn-0	2.4.3	2.4.3
Hadoop	3.3.3-amzn-0	3.2.1-amzn-3	2.8.5-amzn-5	2.8.5-amzn-1
Scala	2.12	2.12	2.11	2.11
Jackson	2.13.3	2.10.x	2.7.x	2.7.x
Hive	2.3.9-amzn-2	2.3.7-amzn-4	1.2	1.2
EMRFS	2.54.0	2.46.0	2.38.0	2.30.0
Json4s	3.7.0-M11	3.6.6	3.5.x	3.5.x
Arrow	7.0.0	2.0.0	0.10.0	0.10.0
AWS Glue Data Catalog client	3.7.0	3.0.0	1.10.0	N/A
Python	3.10	3.7	2.7 & 3.6	2.7 & 3.6
Boto	1.26	1.18	1.12	N/A

Appendix B: JDBC driver upgrades

The following are JDBC driver upgrades:

Driver	JDBC driver version in past AWS Glue versions	JDBC driver version in AWS Glue 3.0	JDBC driver version in AWS Glue 4.0
MySQL	5.1	8.0.23	8.0.23
Microsoft SQL Server	6.1.0	7.0.0	9.4.0
Oracle Databases	11.2	21.1	21.7
PostgreSQL	42.1.0	42.2.18	42.3.6
MongoDB	2.0.0	4.0.0	4.7.2
Amazon Redshift	redshift-jdbc41-1.2.12.1017	redshift-jdbc41-1.2.12.1017	redshift-jdbc42-2.1.0.16

Appendix C: Connector upgrades

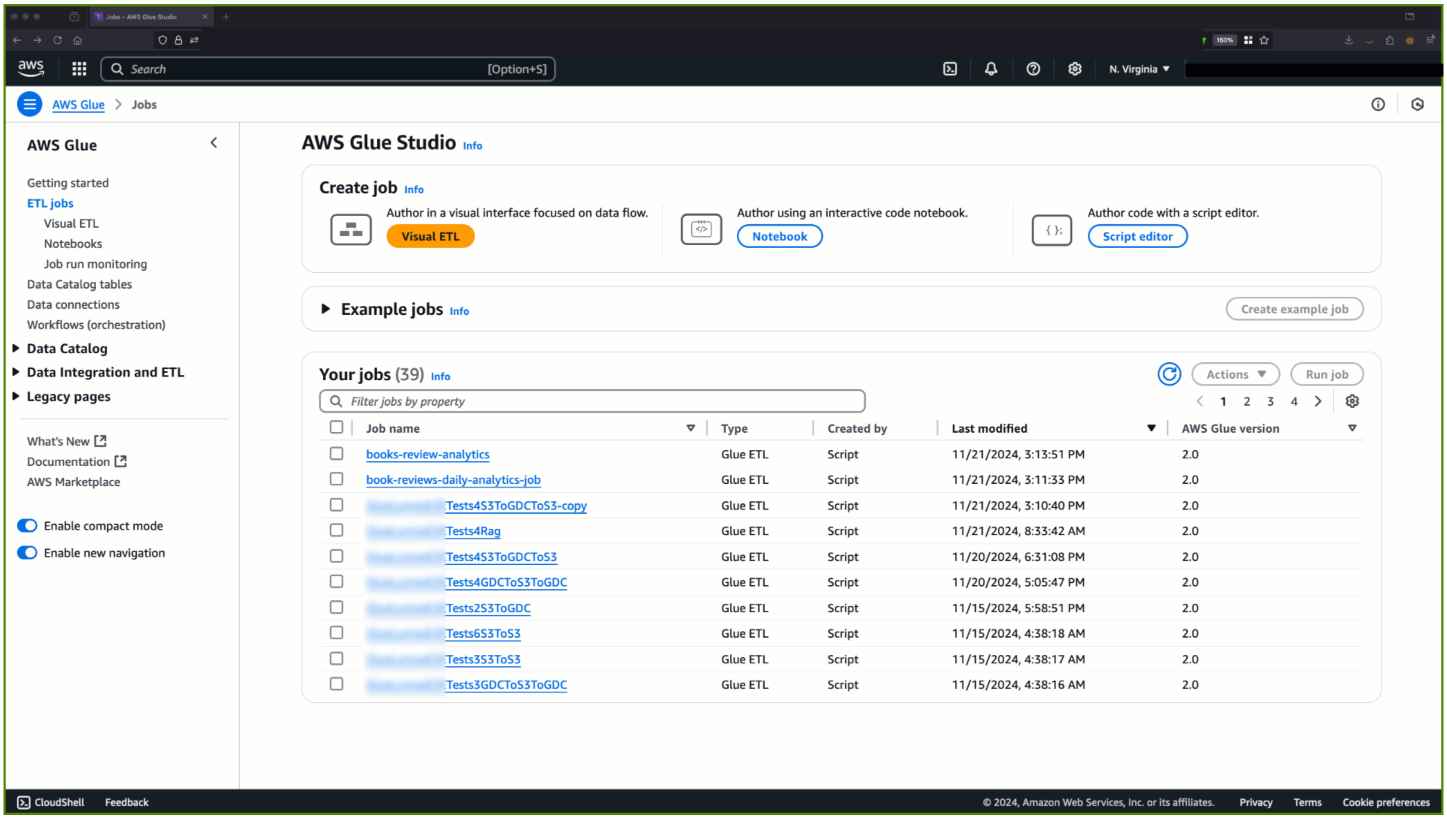
The following are connector upgrades:

Driver	Connector version in AWS Glue 3.0	Connector version in AWS Glue 4.0
MongoDB	3.0.0	10.0.4
Hudi	0.10.1	0.12.1
Delta Lake	1.0.0	2.1.0
Iceberg	0.13.1	1.0.0
DynamoDB	1.11	1.12

Generative AI upgrades for Apache Spark in AWS Glue

The generative AI upgrades for Apache Spark preview is available for AWS Glue versions 4.0 and 5.0 in the following AWS Regions: US East (Ohio), US East (N. Virginia), US West (Oregon), Asia Pacific (Tokyo), and Asia Pacific (Sydney). Preview features are subject to change.

Spark Upgrades in AWS Glue enables data engineers and developers to upgrade and migrate their existing AWS Glue Spark jobs to the latest Spark releases using generative AI. Data engineers can use it to scan their AWS Glue Spark jobs, generate upgrade plans, execute plans, and validate outputs. It reduces the time and cost of Spark upgrades by automating the undifferentiated work of identifying and updating Spark scripts, configurations, dependencies, methods, and features.


[redacted]
AWS Glue > Jobs
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Data Catalog
Data Integration and ETL
Legacy pages
What's New
Documentation
AWS Marketplace
Enable compact mode
Enable new navigation
AWS Glue Studio Info
Create job Info
Author in a visual interface focused on data flow.
Visual ETL
Author using an interactive code notebook.
Notebook
Author code with a script editor.
Script editor
Example jobs Info
Create example job
Your jobs (39) Info
Actions
Run job
Filter jobs by property
1 2 3 4
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

☐	Job name	Type	Created by	Last modified	AWS Glue version
☐	books-review-analytics	Glue ETL	Script	11/21/2024, 3:13:51 PM	2.0
☐	book-reviews-daily-analytics-job	Glue ETL	Script	11/21/2024, 3:11:33 PM	2.0
☐	[redacted]Tests4S3ToGDCToS3-copy	Glue ETL	Script	11/21/2024, 3:10:40 PM	2.0
☐	[redacted]Tests4Rag	Glue ETL	Script	11/21/2024, 8:33:42 AM	2.0
☐	[redacted]Tests4S3ToGDCToS3	Glue ETL	Script	11/20/2024, 6:31:08 PM	2.0
☐	[redacted]Tests4GDCToS3ToGDC	Glue ETL	Script	11/20/2024, 5:05:47 PM	2.0
☐	[redacted]Tests2S3ToGDC	Glue ETL	Script	11/15/2024, 5:58:51 PM	2.0
☐	[redacted]Tests6S3ToS3	Glue ETL	Script	11/15/2024, 4:38:18 AM	2.0
☐	[redacted]Tests3S3ToS3	Glue ETL	Script	11/15/2024, 4:38:17 AM	2.0
☐	[redacted]Tests3GDCToS3ToGDC	Glue ETL	Script	11/15/2024, 4:38:16 AM	2.0

How it works

When you use upgrade analysis, AWS Glue identifies differences between versions and configurations in your job's code to generate an upgrade plan. The upgrade plan details all code changes, and required migration steps. Next, AWS Glue builds and runs the upgraded application in an environment to validate changes and generates a list of code changes for you to migrate your job. You can view the updated script along with the summary that details the proposed changes. After running your own tests, accept the changes and the AWS Glue job will be updated automatically to the latest version with the new script.

The upgrade analysis process can take some time to complete, depending on the complexity of the job and the workload. The results of the upgrade analysis will be stored in the specified Amazon S3 path, which can be reviewed to understand the upgrade and any potential compatibility issues. After reviewing the upgrade analysis results, you can decide whether to proceed with the actual upgrade or make any necessary changes to the job before upgrading.

Prerequisites

The following prerequisites are required to use generative AI to upgrade jobs in AWS Glue:

- AWS Glue 2 PySpark jobs – only AWS Glue 2 jobs can be upgraded to AWS Glue 5.
- IAM permissions are required to start the analysis, review the results and upgrade your job. For more information, see the examples in the Permissions section below.
- If using AWS KMS to encrypt analysis artifacts, then additional AWS AWS KMS permissions are needed. For more information, see the examples in the AWS KMS policy section below.

Permissions

To start a new upgrade analysis, you need the following permissions:

1. Update the IAM policy of the caller with the following permission:

   JSON

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "glue:StartJobUpgradeAnalysis",
                "glue:StartJobRun",
                "glue:GetJobRun",
                "glue:GetJob",
                "glue:BatchStopJobRun"
            ],
            "Resource": [
                "arn:aws:glue:us-east-1:111122223333:job/jobName"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "s3:GetObject"
            ],
            "Resource": [
              "arn:aws:s3:::amzn-s3-demo-bucket/script-location/*"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "s3:PutObject",
                "s3:GetObject"
            ],
            "Resource": [
              "arn:aws:s3:::amzn-s3-demo-bucket/results/*"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "kms:Decrypt",
                "kms:GenerateDataKey"
            ],
            "Resource": "arn:aws:kms:us-east-1:111122223333:key/key-id"
        }
    ]
}
```

2. Update the Execution role of the job you are upgrading to include the following in-line policy:

```
    {
      "Effect": "Allow",
      "Action": ["s3:GetObject"],
      "Resource": [
        "ARN of the Amazon S3 path provided on API",
        "ARN of the Amazon S3 path provided on API/*"
      ]
    }
```

For example, if you are using the Amazon S3 path `s3://amzn-s3-demo-bucket/upgraded-result`, then the policy will be:

```
{
      "Effect": "Allow",
      "Action": ["s3:GetObject"],
      "Resource": [
        "arn:aws:s3:::amzn-s3-demo-bucket/upgraded-result/",
        "arn:aws:s3:::amzn-s3-demo-bucket/upgraded-result/*"
      ]
    }
```

To retrieve the details of an analysis, you need the following permissions:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:GetJobUpgradeAnalysis"
      ],
      "Resource": [
        "arn:aws:glue:us-east-1:111122223333:job/jobName"
      ]
    }
  ]
}
```

To stop an in-progress analysis, you need the following permissions:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:StopJobUpgradeAnalysis",
        "glue:BatchStopJobRun"
      ],
      "Resource": [
        "arn:aws:glue:us-east-1:111122223333:job/jobName"
      ]
    }
  ]
}
```

To list all the analyses submitted on a specific job, you need the following permissions:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListJobUpgradeAnalyses"
      ],
      "Resource": [
        "arn:aws:glue:us-east-1:111122223333:job/jobName"
      ]
    }
  ]
}
```

To accept the changes from an analysis and upgrade a job, you need the following permissions:

AWS KMS policy

To pass your own custom AWS KMS key when starting an analysis, please refer to the following section to configure appropriate permissions on the AWS KMS keys.

Configuring the result artifact encryption using a AWS KMS key:

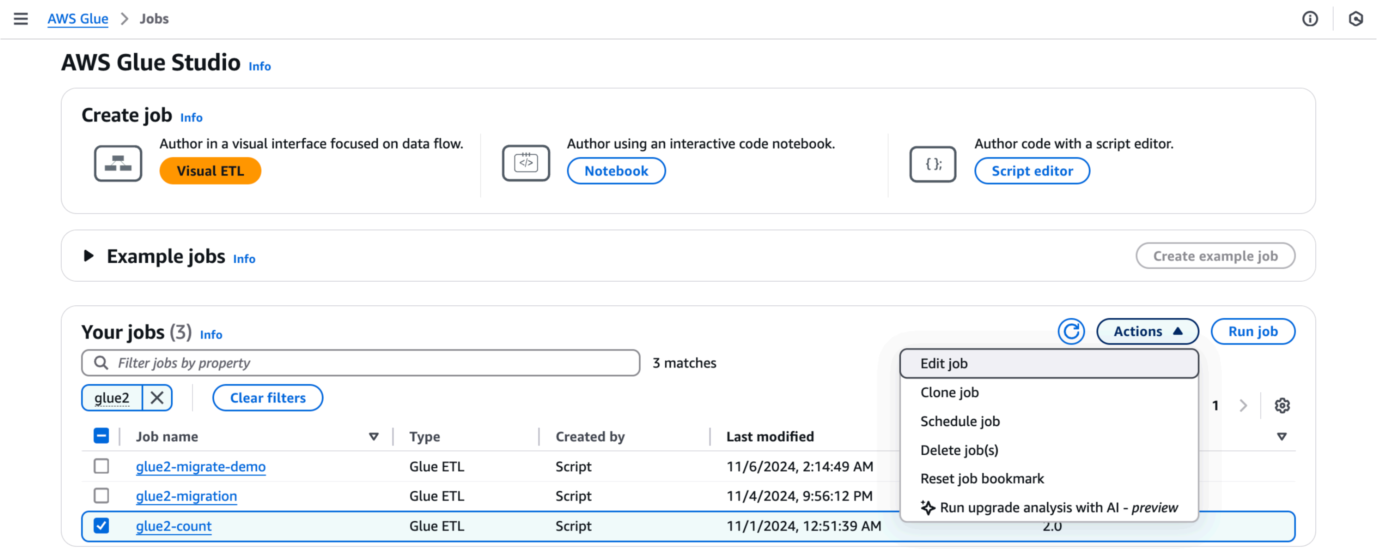
This policy ensures that you have both the encryption and decryption permissions on the AWS KMS key.

```
{
    "Effect": "Allow",
    "Principal":{
        "AWS": "<IAM Customer caller ARN>"
    },
    "Action": [
      "kms:Decrypt",
      "kms:GenerateDataKey",
    ],
    "Resource": "<key-arn-passed-on-start-api>"
}
```

Running an upgrade analysis and applying the upgrade script

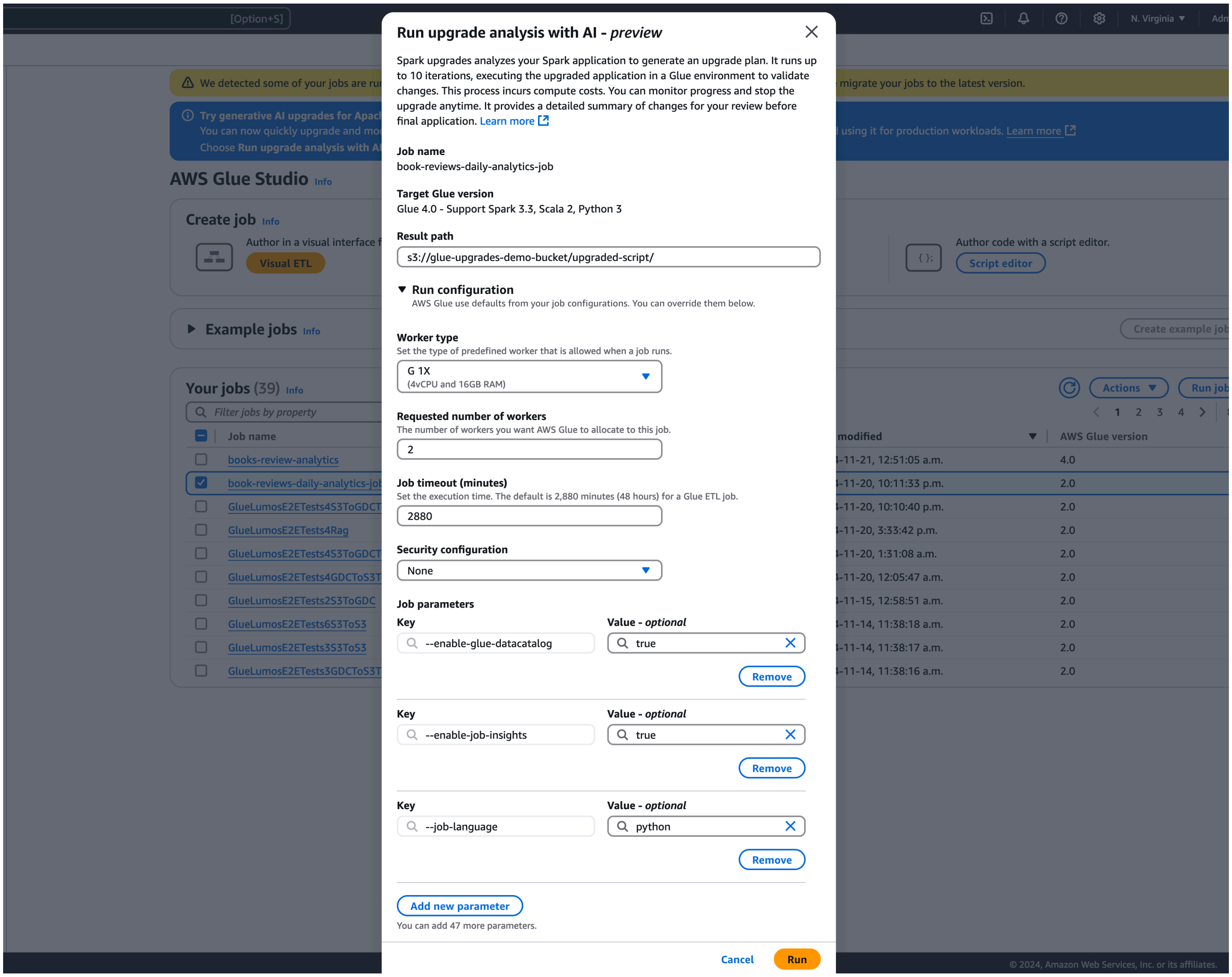
You can run an upgrade analysis, which will generate an upgrade plan on a job you select from the **Jobs** view.

1. From **Jobs**, select a AWS Glue 2.0 job, then choose **Run upgrade analysis** from the **Actions** menu.


AWS Glue > Jobs
AWS Glue Studio Info
Create job Info
Author in a visual interface focused on data flow.
Visual ETL
Author using an interactive code notebook.
Notebook
Author code with a script editor.
Script editor
Example jobs Info
Create example job
Your jobs (3) Info
Actions ▲
Run job
Filter jobs by property
3 matches
glue2
Clear filters
Edit job
Clone job
Schedule job
Delete job(s)
Reset job bookmark
Run upgrade analysis with AI - preview
1

	Job name	Type	Created by	Last modified	
☐	glue2-migrate-demo	Glue ETL	Script	11/6/2024, 2:14:49 AM	
☐	glue2-migration	Glue ETL	Script	11/4/2024, 9:56:12 PM	
☑	glue2-count	Glue ETL	Script	11/1/2024, 12:51:39 AM	2.0

2. In the modal, select a path to store your generated upgrade plan in the **Result path**. This must be an Amazon S3 bucket you can access and write to.



Run upgrade analysis with AI - ***preview***

Spark upgrades analyzes your Spark application to generate an upgrade plan. It runs up to 10 iterations, executing the upgraded application in a Glue environment to validate changes. This process incurs compute costs. You can monitor progress and stop the upgrade anytime. It provides a detailed summary of changes for your review before final application. Learn more

Job name
book-reviews-daily-analytics-job

Target Glue version
Glue 4.0 - Support Spark 3.3, Scala 2, Python 3

Result path
s3://glue-upgrades-demo-bucket/upgraded-script/

Run configuration
AWS Glue use defaults from your job configurations. You can override them below.

Worker type
Set the type of predefined worker that is allowed when a job runs.
G 1X
(4vCPU and 16GB RAM)

Requested number of workers
The number of workers you want AWS Glue to allocate to this job.
2

Job timeout (minutes)
Set the execution time. The default is 2,880 minutes (48 hours) for a Glue ETL job.
2880

Security configuration
None

Job parameters

Key	Value - *optional*
--enable-glue-datacatalog	true
--enable-job-insights	true
--job-language	python

You can add 47 more parameters.

3. Configure additional options, if needed:

- **Run configuration** – optional: The run configuration is an optional setting that allows you to customize various aspects of the validation runs performed during the upgrade analysis. This configuration is utilized to execute the upgraded script and allows you to select the compute environment properties (worker type, number of workers, etc). Note you should use your non-production developer accounts to run the validations on sample datasets before reviewing, accepting the changes and applying them to production environments. The run configuration includes the following customizable parameters:
  - Worker type: You can specify the type of worker to be used for the validation runs, allowing you to choose appropriate compute resources based on your requirements.
  - Number of workers: You can define the number of workers to be provisioned for the validation runs, enabling you to scale resources according to your workload needs.

    - Job timeout (in minutes): This parameter allows you to set a time limit for the validation runs, ensuring that the jobs terminate after a specified duration to prevent excessive resource consumption.
    - Security configuration: You can configure security settings, such as encryption and access control, to ensure the protection of your data and resources during the validation runs.
    - Additional job parameters: If needed, you can add new job parameters to further customize the execution environment for the validation runs.

  By leveraging the run configuration, you can tailor the validation runs to suit your specific requirements. For example, you can configure the validation runs to use a smaller dataset, which allows the analysis to complete more quickly and optimizes costs. This approach ensures that the upgrade analysis is performed efficiently while minimizing resource utilization and associated costs during the validation phase.

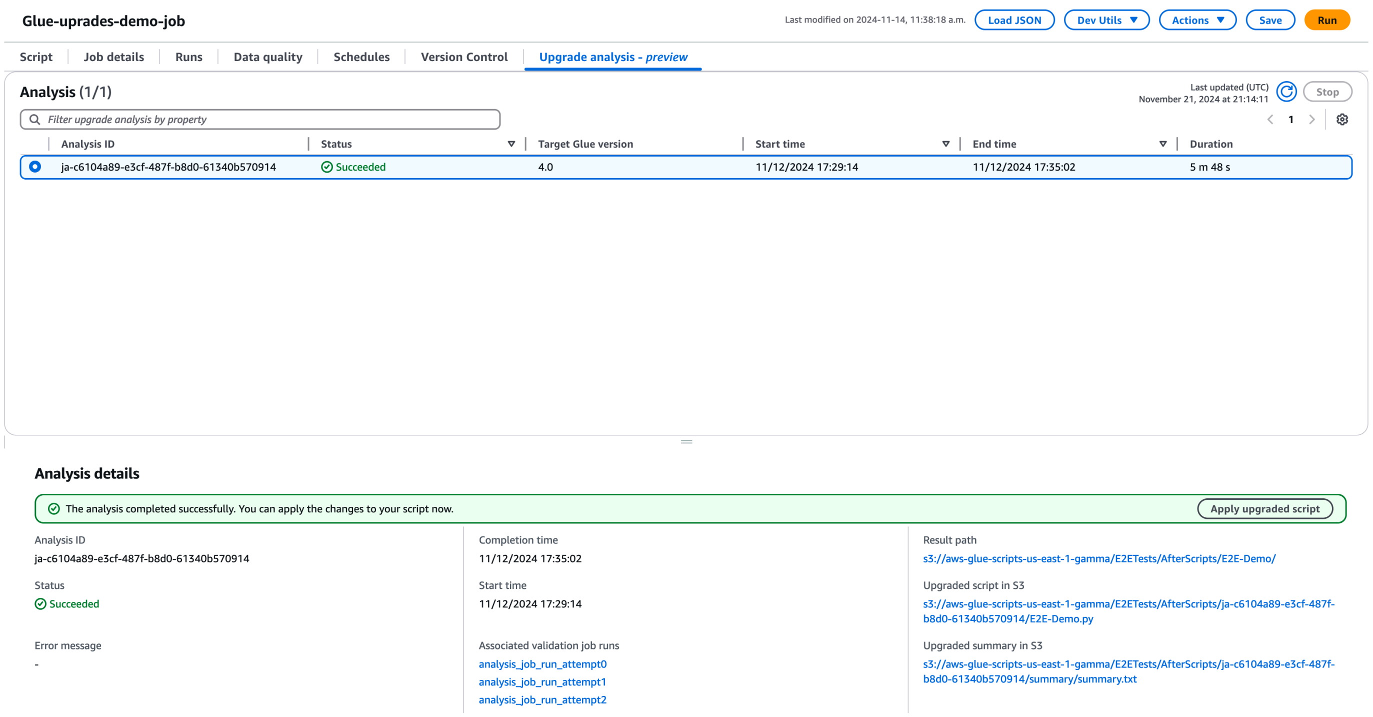
  - **Encryption configuration** – optional:
    - **Enable upgrade artifacts encryption**: Enable at-rest encryption when writing data to the result path. If you don't want to encrypt your upgrade artifacts, leave this option unchecked.

4. Choose **Run** to start the upgrade analysis. While the analysis is running, you can view the results on the **Upgrade analysis** tab. The analysis details window will show you information about the analysis as well as links to the upgrade artifacts.

   - **Result path** – this is where the results summary and upgrade script are stored.
   - **Upgraded script in Amazon S3** – the location of the upgrade script in Amazon S3. You can view the script prior to applying the upgrade.
   - **Upgrade summary in Amazon S3** – the location of the upgrade summary in Amazon S3. You can view the upgrade summary prior to applying the upgrade.

5. When the upgrade analysis is completed successfully, you can apply the upgrade script to automatically upgrade your job by choosing **Apply upgraded script**.

   Once applied, the AWS Glue version will be updated to 4.0. You can view the new script in the **Script** tab.



Glue-uprades-demo-job

Last modified on 2024-11-14, 11:38:18 a.m.

Analysis (1/1)

Last updated (UTC)
November 21, 2024 at 21:14:11

	Analysis ID	Status	Target Glue version	Start time	End time	Duration
●	ja-c6104a89-e3cf-487f-b8d0-61340b570914	Succeeded	4.0	11/12/2024 17:29:14	11/12/2024 17:35:02	5 m 48 s

Analysis details

The analysis completed successfully. You can apply the changes to your script now.

Analysis ID
ja-c6104a89-e3cf-487f-b8d0-61340b570914

Status
Succeeded

Error message
-

Completion time
11/12/2024 17:35:02

Start time
11/12/2024 17:29:14

Associated validation job runs
analysis_job_run_attempt0
analysis_job_run_attempt1
analysis_job_run_attempt2

Result path
s3://aws-glue-scripts-us-east-1-gamma/E2ETests/AfterScripts/E2E-Demo/

Upgraded script in S3
s3://aws-glue-scripts-us-east-1-gamma/E2ETests/AfterScripts/ja-c6104a89-e3cf-487f-b8d0-61340b570914/E2E-Demo.py

Upgraded summary in S3
s3://aws-glue-scripts-us-east-1-gamma/E2ETests/AfterScripts/ja-c6104a89-e3cf-487f-b8d0-61340b570914/summary/summary.txt

Understanding your upgrade summary

This example demonstrates the process of upgrading a AWS Glue job from version 2.0 to version 4.0. The sample job reads product data from an Amazon S3 bucket, applies several transformations to the data using Spark SQL, and then saves the transformed results back to an Amazon S3 bucket.

Original code (AWS Glue 2.0) - before upgrade

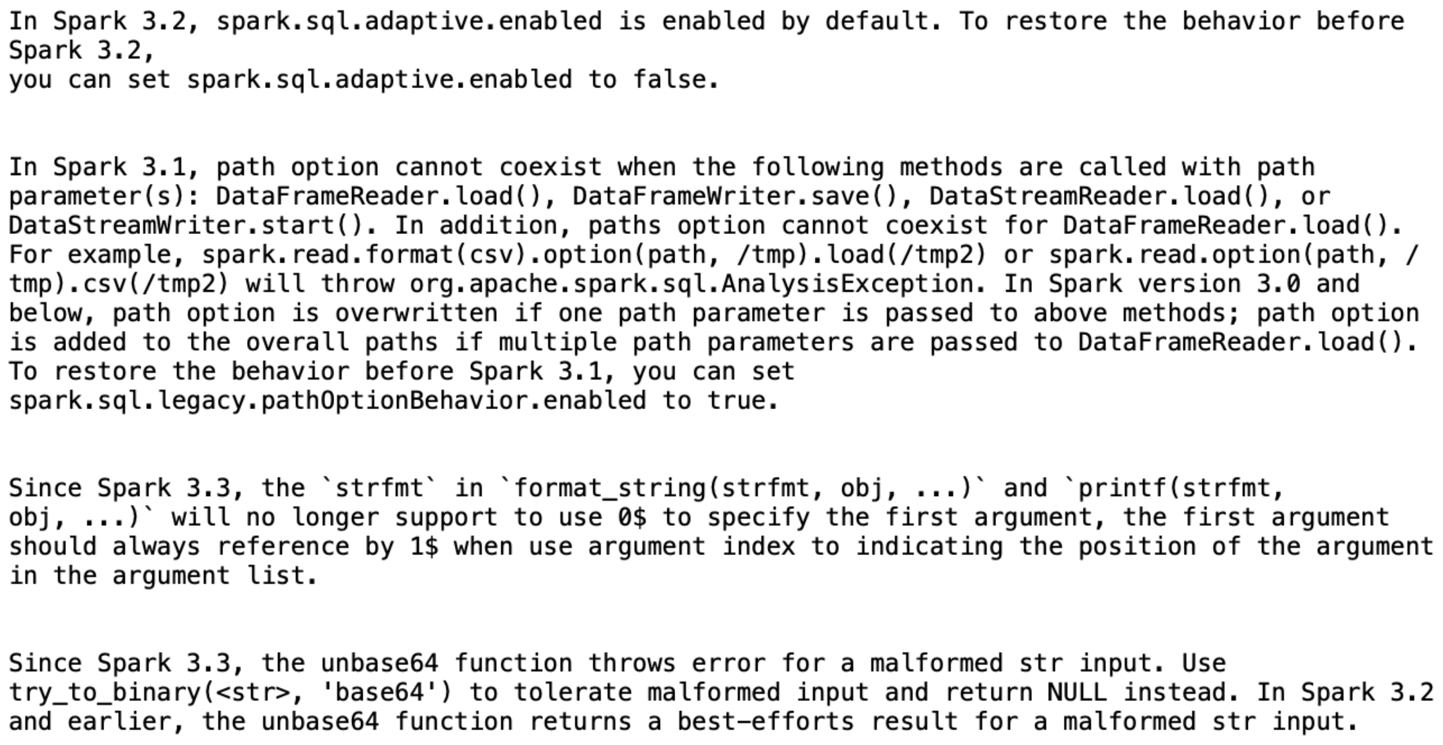
```
from awsglue.transforms import *
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from pyspark.sql.types import *
from pyspark.sql.functions import *
from awsglue.job import Job
import json
from pyspark.sql.types import StructType

sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
```

```
gdc_database = "s3://aws-glue-scripts-us-east-1-gamma/demo-database/"
schema_location = (
    "s3://aws-glue-scripts-us-east-1-gamma/DataFiles/"
)

products_schema_string = spark.read.text(
    f"{schema_location}schemas/products_schema"
).first()[0]

product_schema = StructType.fromJson(json.loads(products_schema_string))

products_source_df = (
    spark.read.option("header", "true")
    .schema(product_schema)
    .option(
        "path",
        f"{gdc_database}products/",
    )
    .csv(f"{gdc_database}products/")
)

products_source_df.show()
products_temp_view_name = "spark_upgrade_demo_product_view"
products_source_df.createOrReplaceTempView(products_temp_view_name)

query = f"select {products_temp_view_name}.*, format_string('%0$s-%0$s', category,
 subcategory) as unique_category from {products_temp_view_name}"
products_with_combination_df = spark.sql(query)
products_with_combination_df.show()

products_with_combination_df.createOrReplaceTempView(products_temp_view_name)
product_df_attribution = spark.sql(
    f"""
SELECT *,
unbase64(split(product_name, ' ')[0]) as product_name_decoded,
unbase64(split(unique_category, '-')[1]) as subcategory_decoded
FROM {products_temp_view_name}
"""
)
product_df_attribution.show()


product_df_attribution.write.mode("overwrite").option("header", "true").option(
    "path", f"{gdc_database}spark_upgrade_demo_product_agg/"
).saveAsTable("spark_upgrade_demo_product_agg", external=True)

spark_upgrade_demo_product_agg_table_df = spark.sql(
    f"SHOW TABLE EXTENDED in default like 'spark_upgrade_demo_product_agg'"
)
spark_upgrade_demo_product_agg_table_df.show()
job.commit()
```

New code (Glue 4.0) - after upgrade

```
from awsglue.transforms import *
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from pyspark.sql.types import *
from pyspark.sql.functions import *
from awsglue.job import Job
import json
from pyspark.sql.types import StructType

sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
# change 1
spark.conf.set("spark.sql.adaptive.enabled", "false")
# change 2
spark.conf.set("spark.sql.legacy.pathOptionBehavior.enabled", "true")
job = Job(glueContext)

gdc_database = "s3://aws-glue-scripts-us-east-1-gamma/demo-database/"
schema_location = (
    "s3://aws-glue-scripts-us-east-1-gamma/DataFiles/"
)

products_schema_string = spark.read.text(
    f"{schema_location}schemas/products_schema"
).first()[0]

product_schema = StructType.fromJson(json.loads(products_schema_string))

products_source_df = (
    spark.read.option("header", "true")
    .schema(product_schema)
    .option(
        "path",
        f"{gdc_database}products/",
    )
    .csv(f"{gdc_database}products/")
)

products_source_df.show()
products_temp_view_name = "spark_upgrade_demo_product_view"
products_source_df.createOrReplaceTempView(products_temp_view_name)

# change 3
query = f"select {products_temp_view_name}.*, format_string('%1$s-%1$s', category,
 subcategory) as unique_category from {products_temp_view_name}"
products_with_combination_df = spark.sql(query)
products_with_combination_df.show()

products_with_combination_df.createOrReplaceTempView(products_temp_view_name)
# change 4
product_df_attribution = spark.sql(
    f"""
SELECT *,
try_to_binary(split(product_name, ' ')[0], 'base64') as product_name_decoded,
try_to_binary(split(unique_category, '-')[1], 'base64') as subcategory_decoded
FROM {products_temp_view_name}
"""
)
product_df_attribution.show()


product_df_attribution.write.mode("overwrite").option("header", "true").option(
    "path", f"{gdc_database}spark_upgrade_demo_product_agg/"
).saveAsTable("spark_upgrade_demo_product_agg", external=True)

spark_upgrade_demo_product_agg_table_df = spark.sql(
    f"SHOW TABLE EXTENDED in default like 'spark_upgrade_demo_product_agg'"
)
spark_upgrade_demo_product_agg_table_df.show()
job.commit()
```

Explanation of analysis summary



```
In Spark 3.2, spark.sql.adaptive.enabled is enabled by default. To restore the behavior before
Spark 3.2,
you can set spark.sql.adaptive.enabled to false.


In Spark 3.1, path option cannot coexist when the following methods are called with path
parameter(s): DataFrameReader.load(), DataFrameWriter.save(), DataStreamReader.load(), or
DataStreamWriter.start(). In addition, paths option cannot coexist for DataFrameReader.load().
For example, spark.read.format(csv).option(path, /tmp).load(/tmp2) or spark.read.option(path, /
tmp).csv(/tmp2) will throw org.apache.spark.sql.AnalysisException. In Spark version 3.0 and
below, path option is overwritten if one path parameter is passed to above methods; path option
is added to the overall paths if multiple path parameters are passed to DataFrameReader.load().
To restore the behavior before Spark 3.1, you can set
spark.sql.legacy.pathOptionBehavior.enabled to true.


Since Spark 3.3, the `strfmt` in `format_string(strfmt, obj, ...)` and `printf(strfmt,
obj, ...)` will no longer support to use 0$ to specify the first argument, the first argument
should always reference by 1$ when use argument index to indicating the position of the argument
in the argument list.


Since Spark 3.3, the unbase64 function throws error for a malformed str input. Use
try_to_binary(<str>, 'base64') to tolerate malformed input and return NULL instead. In Spark 3.2
and earlier, the unbase64 function returns a best-efforts result for a malformed str input.
```

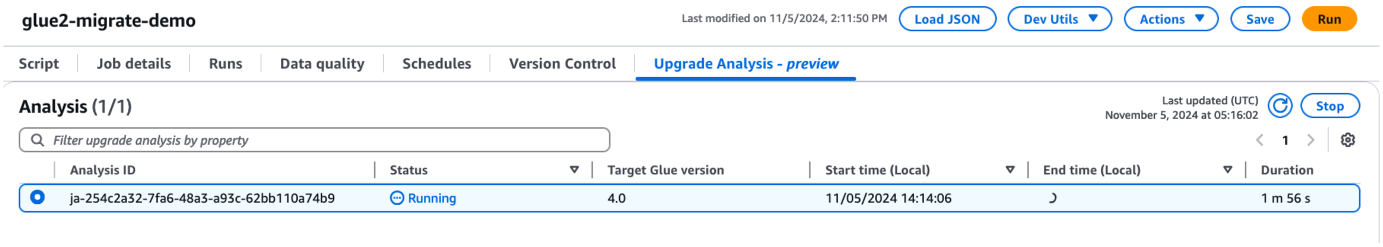
Based on the summary, there are four changes proposed by AWS Glue in order to successfully upgrade the script from AWS Glue 2.0 to AWS Glue 4.0:

1. **Spark SQL configuration (spark.sql.adaptive.enabled)**: This change is to restore the application behavior as a new feature for Spark SQL adaptive query execution is introduced starting Spark 3.2. You can inspect this configuration change and can further enable or disable it as per their preference.
2. **DataFrame API change**: The path option cannot coexist with other DataFrameReader operations like `load()`. To retain the previous behavior, AWS Glue updated the script to add a new SQL configuration **(spark.sql.legacy.pathOptionBehavior.enabled)**.
3. **Spark SQL API change**: The behavior of `strfmt` in `format_string(strfmt, obj, ...)` has been updated to disallow `0$` as the first argument. To ensure compatibility, AWS Glue has modified the script to use `1$` as the first argument instead.
4. **Spark SQL API change**: The `unbase64` function does not allow malformed string inputs. To retain the previous behavior, AWS Glue updated the script to use the `try_to_binary` function.

Stopping an upgrade analysis in progress

You can cancel an upgrade analysis in progress or just stop the analysis.

1. Choose the **Upgrade Analysis** tab.
2. Select the job that is running, then choose **Stop**. This will stop the analysis. You can then run another upgrade analysis on the same job.



glue2-migrate-demo

Last modified on 11/5/2024, 2:11:50 PM

Analysis (1/1)

Last updated (UTC) November 5, 2024 at 05:16:02

	Analysis ID	Status	Target Glue version	Start time (Local)	End time (Local)	Duration
●	ja-254c2a32-7fa6-48a3-a93c-62bb110a74b9	Running	4.0	11/05/2024 14:14:06		1 m 56 s

Considerations

As you begin using Spark Upgrades during the preview period, there are several important aspects to consider for optimal usage of the service.

- **Service Scope and Limitations**: The preview release focuses on PySpark code upgrades from AWS Glue versions 2.0 to version 5.0. At this time, the service handles PySpark code that doesn't rely on additional library dependencies. You can run automated upgrades for up to 10 jobs concurrently in an AWS account, allowing you to efficiently upgrade multiple jobs while maintaining system stability.
    - Only PySpark jobs are supported.
    - Upgrade analysis will time out after 24 hours.
    - Only one active upgrade analysis can be run at a time for one job. On the account-level, up to 10 active upgrade analysis can be run at the same time.
- **Optimizing Costs During Upgrade Process**: Since Spark Upgrades use generative AI to validate the upgrade plan through multiple iterations, with each iteration running as a AWS Glue job in your account, it’s essential to optimize the validation job run configurations for cost efficiency. To achieve this, we recommend specifying a Run Configuration when starting an Upgrade Analysis as follows:
    - Use non-production developer accounts and select sample mock datasets that represent your production data but are smaller in size for validation with Spark Upgrades.

  - Using right-sized compute resources, such as G.1X workers, and selecting an appropriate number of workers for processing your sample data.
  - Enabling AWS Glue job auto-scaling when applicable to automatically adjust resources based on workload.

  For example, if your production job processes terabytes of data with 20 G.2X workers, you might configure the upgrade job to process a few gigabytes of representative data with 2 G.2X workers and auto-scaling enabled for validation.

- **Preview Best Practices**: During the preview period, we strongly recommend starting your upgrade journey with non-production jobs. This approach allows you to familiarize yourself with the upgrade workflow, and understand how the service handles different types of Spark code patterns.
- **Alarms and notifications**: When utilizing the Generative AI upgrades feature on a job, ensure that alarms/notifications for failed job runs are turned off. During the upgrade process, there may be up to 10 failed job runs in your account before the upgraded artifacts are provided.
- **Anomaly detection rules**: Turn off any anomaly detection rules on the Job that is being upgraded as well, as the data written to output folders during intermediate job runs might not be in the expected format while the upgrade validation is in progress.
- **Use upgrade analysis with idempotent jobs**: Use upgrade analysis with idempotent jobs to ensure each subsequent validation job run attempt is similar to the previous one, and doesn't run into issues. Idempotent jobs are jobs that can be run multiple times with the same input data, and they will produce the same output each time. When using the Generative AI upgrades for Apache Spark in AWS Glue, the service will run multiple iterations of your job as part of the validation process. During each iteration, it will make changes to your Spark code and configurations to validate the upgrade plan. If your Spark job is not idempotent, running it multiple times with the same input data could lead to issues.

Cross-region inference in Spark Upgrades

Spark Upgrades is powered by Amazon Bedrock and leverages cross-region inference (CRIS). With CRIS, Spark Upgrades will automatically select the optimal region within your geography (as described in more detail here) to process your inference request, maximizing available compute resources and model availability, and providing the best customer experience. There's no additional cost for using cross-region inference.

Cross-region inference requests are kept within the AWS Regions that are part of the geography where the data originally resides. For example, a request made within the US is kept within the AWS Regions in the US. Although the data remains stored only in the primary region, when using cross-region inference, your input prompts and output results may move outside of your primary region. All data will be transmitted encrypted across Amazon's secure network.

Working with Spark jobs in AWS Glue

Provides information on AWS Glue for Spark ETL jobs.

Topics

- Using job parameters in AWS Glue jobs
- AWS Glue Spark and PySpark jobs
- AWS Glue worker types
- Streaming ETL jobs in AWS Glue
- Record matching with AWS Lake Formation FindMatches
- Migrate Apache Spark programs to AWS Glue

Using job parameters in AWS Glue jobs

When creating a AWS Glue job, you set some standard fields, such as `Role` and `WorkerType`. You can provide additional configuration information through the `Argument` fields (**Job Parameters** in the console). In these fields, you can provide AWS Glue jobs with the arguments (parameters) listed in this topic.

For more information about the AWS Glue Job API, see the section called “Jobs”.

Note

Job arguments have a maximum size limit of 260KB. A validation check will raise an error if the argument size is greater than 260KB.

Setting job parameters

You can configure a job through the console on the **Job details** tab, under the **Job Parameters** heading. You can also configure a job through the AWS CLI by setting `DefaultArguments` or `NonOverridableArguments` on a job, or setting `Arguments` on a job run. Arguments set on the job will be passed in every time the job is run, while arguments set on the job run will only be passed in for that individual run.

For example, the following is the syntax for running a job using `--arguments` to set a job parameter.

```
$ aws glue start-job-run --job-name "CSV to CSV" --arguments='--scriptLocation="s3://
my_glue/libraries/test_lib.py"'
```

Accessing job parameters

When writing AWS Glue scripts, you may want to access job parameter values to alter the behavior of your own code. We provide helper methods to do so in our libraries. These methods resolve job run parameter values that override job parameter values. When resolving parameters set in multiple places, job `NonOverridableArguments` will override job run `Arguments`, which will override job `DefaultArguments`.

In Python:

In Python jobs, we provide a function named `getResolvedParameters`. For more information, see the section called "getResolvedOptions". Job parameters are available in the `sys.argv` variable.

In Scala:

In Scala jobs, we provide an object named `GlueArgParser`. For more information, see the section called "GlueArgParser". Job parameters are available in the `sysArgs` variable.

Job parameter reference

AWS Glue recognizes the following argument names that you can use to set up the script environment for your jobs and job runs:

`--additional-python-modules`

A comma delimited list representing a set of Python packages to be installed. You can install packages from PyPI or provide a custom distribution. A PyPI package entry will be in the format *`package`*`==`*`version`*, with the PyPI name and version of your target package. A custom distribution entry is the S3 path to the distribution.

Entries use Python version matching to match package and version. This means you will need to use two equals signs, such as ==. There are other version matching operators, for more information see PEP 440.

To pass module installation options to `pip3`, use the --python-modules-installer-option parameter.

`--auto-scale-within-microbatch`

The default value is true. This parameter can only be used for AWS Glue streaming jobs, which process the streaming data in a series of micro batches, and auto scaling must be enabled. When setting this value to false, it computes the exponential moving average of batch duration for completed micro-batches and compares this value with the window size to determine whether to scale up or scale down the number of executors. Scaling only happens when a micro batch is completed. When setting this value to true, during a micro-batch, it scales up when the number of Spark tasks remains the same for 30 seconds, or the current batch processing is greater than the window size. The number of executors will drop if an executor has been idle for more than 60 seconds, or the exponential moving average of batch duration is low.

`--class`

The Scala class that serves as the entry point for your Scala script. This applies only if your `--job-language` is set to `scala`.

`--continuous-log-conversionPattern`

Specifies a custom conversion log pattern for a job enabled for continuous logging. The conversion pattern applies only to driver logs and executor logs. It does not affect the AWS Glue progress bar.

`--continuous-log-logGroup`

Specifies a custom Amazon CloudWatch log group name for a job enabled for continuous logging.

`--continuous-log-logStreamPrefix`

Specifies a custom CloudWatch log stream prefix for a job enabled for continuous logging.

`--customer-driver-env-vars` and `--customer-executor-env-vars`

These parameters set environment variables on the operating system respectively for each worker (driver or executor). You can use these parameters when building platforms and custom frameworks on top of AWS Glue, to let your users write jobs on top of it. Enabling these two flags will allow you to set different environment variables on the driver and executor respectively without having to inject the same logic in the job script itself.

Example usage

The following is an example of using these parameters:

```
"—customer-driver-env-vars", "CUSTOMER_KEY1=VAL1,CUSTOMER_KEY2=\"val2,val2 val2\"",
"—customer-executor-env-vars", "CUSTOMER_KEY3=VAL3,KEY4=VAL4"
```

Setting these in the job run argument is equivalent to running the following commands:

In the driver:

- export CUSTOMER_KEY1=VAL1
- export CUSTOMER_KEY2="val2,val2 val2"

In the executor:

- export CUSTOMER_KEY3=VAL3

Then, in the job script itself, you can retrieve the environment variables using `os.environ.get("CUSTOMER_KEY1")` or `System.getenv("CUSTOMER_KEY1")`.

Enforced syntax

Observe the following standards when defining environment variables:

- Each key must have the `CUSTOMER_ prefix`.

  For example: for `"CUSTOMER_KEY3=VAL3,KEY4=VAL4"`, `KEY4=VAL4` will be ignored and not set.
- Each key and value pair must be delineated with a single comma.

  For example: `"CUSTOMER_KEY3=VAL3,CUSTOMER_KEY4=VAL4"`

- If the "value" has spaces or commas, then it must be defined within quotations.

  For example: `CUSTOMER_KEY2=\"val2,val2 val2\"`

This syntax closely models the standards of setting bash environment variables.

`--datalake-formats`

Supported in AWS Glue 3.0 and later versions.

Specifies the data lake framework to use. AWS Glue adds the required JAR files for the frameworks that you specify into the `classpath`. For more information, see Using data lake frameworks with AWS Glue ETL jobs.

You can specify one or more of the following values, separated by a comma:

- `hudi`
- `delta`
- `iceberg`

For example, pass the following argument to specify all three frameworks.

```
'--datalake-formats': 'hudi,delta,iceberg'
```

`--disable-proxy-v2`

Disable the service proxy to allow AWS service calls to Amazon S3, CloudWatch, and AWS Glue originating from your script through your VPC. For more information, see Configuring AWS calls to go through your VPC . To disable the service proxy, set the value of this paramater to `true`.

`--enable-auto-scaling`

Turns on auto scaling and per-worker billing when you set the value to `true`.

`--enable-continuous-cloudwatch-log`

Enables real-time continuous logging for AWS Glue jobs. You can view real-time Apache Spark job logs in CloudWatch.

`--enable-continuous-log-filter`

Specifies a standard filter (`true`) or no filter (`false`) when you create or edit a job enabled for continuous logging. Choosing the standard filter prunes out non-useful Apache Spark driver/

executor and Apache Hadoop YARN heartbeat log messages. Choosing no filter gives you all the log messages.

`--enable-glue-datacatalog`

Enables you to use the AWS Glue Data Catalog as an Apache Spark Hive metastore. To enable this feature, set the value to `true`.

`--enable-job-insights`

Enables additional error analysis monitoring with AWS Glue job run insights. For details, see the section called “Monitoring with AWS Glue job run insights”. By default, the value is set to `true` and job run insights are enabled.

This option is available for AWS Glue version 2.0 and 3.0.

`--enable-lakeformation-fine-grained-access`

Enables fine-grained access control for AWS Glue jobs. For more information, see the section called “Lake Formation for FGAC”.

`--enable-metrics`

Enables the collection of metrics for job profiling for this job run. These metrics are available on the AWS Glue console and the Amazon CloudWatch console. The value of this parameter is not relevant. To enable this feature, you can provide this parameter with any value, but `true` is recommended for clarity. To disable this feature, remove this parameter from your job configuration.

`--enable-observability-metrics`

Enables a set of Observability metrics to generate insights into what is happening inside each job run on Job Runs Monitoring page under AWS Glue console and the Amazon CloudWatch console. To enable this feature, set the value of this parameter to true. To disable this feature, set it to `false` or remove this parameter from your job configuration.

`--enable-rename-algorithm-v2`

Sets the EMRFS rename algorithm version to version 2. When a Spark job uses dynamic partition overwrite mode, there is a possibility that a duplicate partition is created. For instance, you can end up with a duplicate partition such as `s3://bucket/table/location/p1=1/p1=1`. Here, P1 is the partition that is being overwritten. Rename algorithm version 2 fixes this issue.

This option is only available on AWS Glue version 1.0.

`--enable-s3-parquet-optimized-committer`

Enables the EMRFS S3-optimized committer for writing Parquet data into Amazon S3. You can supply the parameter/value pair via the AWS Glue console when creating or updating an AWS Glue job. Setting the value to **`true`** enables the committer. By default, the flag is turned on in AWS Glue 3.0 and off in AWS Glue 2.0.

For more information, see Using the EMRFS S3-optimized Committer.

`--enable-spark-ui`

When set to `true`, turns on the feature to use the Spark UI to monitor and debug AWS Glue ETL jobs.

`--executor-cores`

Number of spark tasks that can run in parallel. This option is supported on AWS Glue 3.0+. The value should not exceed 2x the number of vCPUs on the worker type, which is 8 on `G.1X`, 16 on `G.2X`, 32 on `G.4X`, 64 on `G.8X`, 96 on `G.12X`, 128 on `G.16X`, and 8 on `R.1X`, 16 on `R.2X`, 32 on `R.4X`, 64 on `R.8X`. You should exercise caution while updating this configuration as it could impact job performance because increased task parallelism causes memory, disk pressure as well as it could throttle the source and target systems (for example: it would cause more concurrent connections on Amazon RDS).

`--extra-files`

The Amazon S3 paths to additional files, such as configuration files that AWS Glue copies to the working directory of your script on the driver node before running it. Multiple values must be complete paths separated by a comma (`,`). The value can be individual files or directory locations. This option is not supported for Python Shell job types.

`--extra-jars`

The Amazon S3 paths to additional files that AWS Glue copies to the driver and executors. AWS Glue also adds these files to the Java classpath before executing your script. Multiple values must be complete paths separated by a comma (`,`). The extension need not be `.jar`

`--extra-py-files`

The Amazon S3 paths to additional Python modules that AWS Glue adds to the Python path on the driver node before running your script. Multiple values must be complete paths separated by a comma (`,`). Only individual files are supported, not a directory path.

`--job-bookmark-option`

Controls the behavior of a job bookmark. The following option values can be set.

--job-bookmark-option value	Description
`job-bookmark-enable`	Keep track of previously processed data. When a job runs, process new data since the last checkpoint.
`job-bookmark-disable`	Always process the entire dataset. You are responsible for managing the output from previous job runs.
`job-bookmark-pause`	Process incremental data since the last successful run or the data in the range identified by the following suboptions, without updating the state of the last bookmark. You are responsible for managing the output from previous job runs. The two suboptions are as follows: • **`job-bookmark-from`** `<from-value>` is the run ID that represents all the input that was processed until the last successful run before and including the specified run ID. The corresponding input is ignored. • **`job-bookmark-to`** `<to-value>` is the run ID that represents all the input that was processed until the last successful run before and including the specified run ID. The corresponding input excluding the input identified by the `<from-value>` is processed by the job. Any input later than this input is also excluded for processing. The job bookmark state is not updated when this option set is specified. The suboptions are optional. However, when used, both suboptions must be provided.

For example, to enable a job bookmark, pass the following argument.

```
'--job-bookmark-option': 'job-bookmark-enable'
```

`--job-language`

The script programming language. This value must be either `scala` or `python`. If this parameter is not present, the default is `python`.

`--python-modules-installer-option`

A plaintext string that defines options to be passed to `pip3` when installing modules with --additional-python-modules. Provide options as you would in the command line, separated by spaces and prefixed by dashes. For more information about usage, see the section called "Installing additional Python modules with pip in AWS Glue 2.0 or later".

> **Note**
>
> This option is not supported for AWS Glue jobs when you use Python 3.9.

`--scriptLocation`

The Amazon Simple Storage Service (Amazon S3) location where your ETL script is located (in the form `s3://path/to/my/script.py`). This parameter overrides a script location set in the `JobCommand` object.

`--spark-event-logs-path`

Specifies an Amazon S3 path. When using the Spark UI monitoring feature, AWS Glue flushes the Spark event logs to this Amazon S3 path every 30 seconds to a bucket that can be used as a temporary directory for storing Spark UI events.

`--TempDir`

Specifies an Amazon S3 path to a bucket that can be used as a temporary directory for the job.

For example, to set a temporary directory, pass the following argument.

```
'--TempDir': 's3-path-to-directory'
```

> **Note**
>
> AWS Glue creates a temporary bucket for jobs if a bucket doesn't already exist in a Region. This bucket might permit public access. You can either modify the bucket in Amazon S3 to set the public access block, or delete the bucket later after all jobs in that Region have completed.

`--use-postgres-driver`

When setting this value to `true`, it prioritizes the Postgres JDBC driver in the class path to avoid a conflict with the Amazon Redshift JDBC driver. This option is only available in AWS Glue version 2.0.

`--user-jars-first`

When setting this value to `true`, it prioritizes the customer's extra JAR files in the classpath. This option is only available in AWS Glue version 2.0 or later.

`--conf`

Controls Spark config parameters. It is for advanced use cases.

`--encryption-type`

Legacy parameter. The corresponding behavior should be configured using security configurations. for more information about security configurations, see the section called “Encrypting data written by AWS Glue”.

AWS Glue uses the following arguments internally and you should never use them:

- `--debug` — Internal to AWS Glue. Do not set.
- `--mode` — Internal to AWS Glue. Do not set.
- `--JOB_NAME` — Internal to AWS Glue. Do not set.
- `--endpoint` — Internal to AWS Glue. Do not set.

AWS Glue supports bootstrapping an environment with Python's `site` module using `sitecustomize` to perform site-specific customizations. Bootstrapping your own initilization functions is recommended for advanced use cases only and is supported on a best-effort basis on AWS Glue 4.0.

The environment variable prefix, `GLUE_CUSTOMER`, is reserved for customer use.

AWS Glue Spark and PySpark jobs

AWS Glue support Spark and PySpark jobs. A Spark job is run in an Apache Spark environment managed by AWS Glue. It processes data in batches. A streaming ETL job is similar to a Spark job, except that it performs ETL on data streams. It uses the Apache Spark Structured Streaming framework. Some Spark job features are not available to streaming ETL jobs.

The following sections provide information on AWS Glue Spark and PySpark jobs.

Topics

- Configuring job properties for Spark jobs in AWS Glue
- Editing Spark scripts in the AWS Glue console
- Jobs (legacy)
- Tracking processed data using job bookmarks
- Storing Spark shuffle data
- Monitoring AWS Glue Spark jobs
- Generative AI troubleshooting for Apache Spark in AWS Glue

Configuring job properties for Spark jobs in AWS Glue

When you define your job on the AWS Glue console, you provide values for properties to control the AWS Glue runtime environment.

Defining job properties for Spark jobs

The following list describes the properties of a Spark job. For the properties of a Python shell job, see Defining job properties for Python shell jobs. For properties of a streaming ETL job, see the section called “Defining job properties for a streaming ETL job”.

The properties are listed in the order in which they appear on the **Add job** wizard on AWS Glue console.

Name

Provide a UTF-8 string with a maximum length of 255 characters.

Description

Provide an optional description of up to 2048 characters.

IAM Role

Specify the IAM role that is used for authorization to resources used to run the job and access data stores. For more information about permissions for running jobs in AWS Glue, see Identity and access management for AWS Glue.

Type

The type of ETL job. This is set automatically based on the type of data sources you select.

- **Spark** runs an Apache Spark ETL script with the job command `glueetl`.
- **Spark Streaming** runs a Apache Spark streaming ETL script with the job command `gluestreaming`. For more information, see the section called “Streaming ETL jobs”.
- **Python shell** run a Python script with the job command `pythonshell`. For more information, see Configuring job properties for Python shell jobs in AWS Glue.

AWS Glue version

AWS Glue version determines the versions of Apache Spark and Python that are available to the job, as specified in the following table.

AWS Glue version	Supported Spark and Python versions
5.0	• Spark 3.5.4 • Python 3.11
4.0	• Spark 3.3.0 • Python 3.10
3.0	• Spark 3.1.1 • Python 3.7

Language

The code in the ETL script defines your job's logic. The script can be coded in Python or Scala. You can choose whether the script that the job runs is generated by AWS Glue or provided by you. You provide the script name and location in Amazon Simple Storage Service (Amazon S3). Confirm that there isn't a file with the same name as the script directory in the path. To learn more about writing scripts, see AWS Glue programming guide.

Worker type

The following worker types are available:

The resources available on AWS Glue workers are measured in DPUs. A DPU is a relative measure of processing power that consists of 4 vCPUs of compute capacity and 16 GB of memory.

- **`G.025X`** – When you choose this type, you also provide a value for **Number of workers**. Each worker maps to 0.25 DPU (2 vCPUs, 4 GB of memory) with 84GB disk (approximately 34GB free). We recommend this worker type for low volume streaming jobs. This worker type is only available for AWS Glue version 3.0 or later streaming jobs.
- **`G.1X`** – When you choose this type, you also provide a value for **Number of workers**. Each worker maps to 1 DPU (4 vCPUs, 16 GB of memory) with 94GB disk (approximately 44GB free). We recommend this worker type for workloads such as data transforms, joins, and queries, to offers a scalable and cost effective way to run most jobs.
- **`G.2X`** – When you choose this type, you also provide a value for **Number of workers**. Each worker maps to 2 DPU (8 vCPUs, 32 GB of memory) with 138GB disk (approximately 78GB free). We recommend this worker type for workloads such as data transforms, joins, and queries, to offers a scalable and cost effective way to run most jobs.
- **`G.4X`** – When you choose this type, you also provide a value for **Number of workers**. Each worker maps to 4 DPU (16 vCPUs, 64 GB of memory) with 256GB disk (approximately 230GB free). We recommend this worker type for jobs whose workloads contain your most demanding transforms, aggregations, joins, and queries.
- **`G.8X`** – When you choose this type, you also provide a value for **Number of workers**. Each worker maps to 8 DPU (32 vCPUs, 128 GB of memory) with 512GB disk (approximately 485GB free). We recommend this worker type for jobs whose workloads contain your most demanding transforms, aggregations, joins, and queries.
- **`G.12X`** – When you choose this type, you also provide a value for **Number of workers**. Each worker maps to 12 DPU (48 vCPUs, 192 GB of memory) with 768GB disk (approximately

741GB free). We recommend this worker type for jobs with very large and resource-intensive workloads that require significant compute capacity.

- **`G.16X` –** When you choose this type, you also provide a value for **Number of workers**. Each worker maps to 16 DPU (64 vCPUs, 256 GB of memory) with 1024GB disk (approximately 996GB free). We recommend this worker type for jobs with the largest and most resource-intensive workloads that require maximum compute capacity.
- **`R.1X` –** When you choose this type, you also provide a value for **Number of workers**. Each worker maps to 1 DPU with memory-optimized configuration. We recommend this worker type for memory-intensive workloads that frequently encounter out-of-memory errors or require high memory-to-CPU ratios.
- **`R.2X` –** When you choose this type, you also provide a value for **Number of workers**. Each worker maps to 2 DPU with memory-optimized configuration. We recommend this worker type for memory-intensive workloads that frequently encounter out-of-memory errors or require high memory-to-CPU ratios.
- **`R.4X` –** When you choose this type, you also provide a value for **Number of workers**. Each worker maps to 4 DPU with memory-optimized configuration. We recommend this worker type for large memory-intensive workloads that frequently encounter out-of-memory errors or require high memory-to-CPU ratios.
- **`R.8X` –** When you choose this type, you also provide a value for **Number of workers**. Each worker maps to 8 DPU with memory-optimized configuration. We recommend this worker type for very large memory-intensive workloads that frequently encounter out-of-memory errors or require high memory-to-CPU ratios.

Worker Type Specifications

The following table provides detailed specifications for all available G worker types:

G Worker Type Specifications

Worker Type	DPU per Node	vCPU	Memory (GB)	Disk (GB)	Free Disk Space (GB)	Spark Executors per Node
G.025X	0.25	2	4	84	34	1
G.1X	1	4	16	94	44	1
G.2X	2	8	32	138	78	1

Worker Type	DPU per Node	vCPU	Memory (GB)	Disk (GB)	Free Disk Space (GB)	Spark Executors per Node
G.4X	4	16	64	256	230	1
G.8X	8	32	128	512	485	1
G.12X	**12**	**48**	**192**	**768**	**741**	**1**
G.16X	**16**	**64**	**256**	**1024**	**996**	**1**

Important: G.12X and G.16X worker types, as well as all R worker types (R.1X through R.8X), have higher startup latency.

You are charged an hourly rate based on the number of DPUs used to run your ETL jobs. For more information, see the AWS Glue pricing page.

For AWS Glue version 1.0 or earlier jobs, when you configure a job using the console and specify a **Worker type** of **Standard**, the **Maximum capacity** is set and the **Number of workers** becomes the value of **Maximum capacity** - 1. If you use the AWS Command Line Interface (AWS CLI) or AWS SDK, you can specify the **Max capacity** parameter, or you can specify both **Worker type** and the **Number of workers**.

For AWS Glue version 2.0 or later jobs, you cannot specify a **Maximum capacity**. Instead, you should specify a **Worker type** and the **Number of workers**.

`G.4X` and **`G.8X`** worker types are available only for AWS Glue version 3.0 or later Spark ETL jobs in the following AWS Regions: US East (Ohio), US East (N. Virginia), US West (N. California), US West (Oregon), Asia Pacific (Mumbai), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Canada (Central), Europe (Frankfurt), Europe (Ireland), Europe (London), Europe (Spain), Europe (Stockholm), and South America (São Paulo).

`G.12X`, **`G.16X`**, and **`R.1X`** through **`R.8X`** worker types are available only for AWS Glue version 4.0 or later Spark ETL jobs in the following AWS Regions: US East (N. Virginia), US West (Oregon), US East (Ohio), Europe (Ireland), and Europe (Frankfurt). Additional regions will be supported in future releases.

Requested number of workers

For most worker types, you must specify the number of workers that are allocated when the job runs.

Job bookmark

Specify how AWS Glue processes state information when the job runs. You can have it remember previously processed data, update state information, or ignore state information. For more information, see the section called “Tracking processed data using job bookmarks”.

Job run queuing

Specifies whether job runs are queued to run later when they cannot run immediately due to service quotas.

When checked, job run queuing is enabled for the job runs. If not populated, the job runs will not be considered for queueing.

If this setting does not match the value set in the job run, then the value from the job run field will be used.

Flex execution

When you configure a job using AWS Studio or the API you may specify a standard or flexible job execution class. Your jobs may have varying degrees of priority and time sensitivity. The standard execution-class is ideal for time-sensitive workloads that require fast job startup and dedicated resources.

The flexible execution class is appropriate for non-urgent jobs such as pre-production jobs, testing, and one-time data loads. Flexible job runs are supported for jobs using AWS Glue version 3.0 or later and `G.1X` or `G.2X` worker types. The new worker types (`G.12X`, `G.16X`, and `R.1X` through `R.8X`) do not support flexible execution.

Flex job runs are billed based on the number of workers running at any point in time. Number of workers may be added or removed for a running flexible job run. Instead of billing as a simple calculation of `Max Capacity` * `Execution Time`, each worker will contribute for the time it ran during the job run. The bill is the sum of (`Number of DPUs per worker` * `time each worker ran`).

For more information, see the help panel in AWS Studio, or Jobs and Job runs.

Number of retries

Specify the number of times, from 0 to 10, that AWS Glue should automatically restart the job if it fails. Jobs that reach the timeout limit are not restarted.

Job timeout

Sets the maximum execution time in minutes. The maximum is 7 days or 10,080 minutes. Otherwise, the jobs will throw an exception.

When the value is left blank, the timeout is defaulted to 2880 minutes.

Any existing AWS Glue jobs that had a timeout value greater than 7 days will be defaulted to 7 days. For instance if you specified a timeout of 20 days for a batch job, it will be stopped on the 7th day.

> **Best practices for job timeouts**
>
> Jobs are billed based on execution time. To avoid unexpected charges, configure timeout values appropriate for the expected execution time of your job.

Advanced Properties

Script filename

A unique script name for your job. Cannot be named **Untitled job**.

Script path

The Amazon S3 location of the script. The path must be in the form `s3://bucket/prefix/path/`. It must end with a slash (/) and not include any files.

Job metrics

Turn on or turn off the creation of Amazon CloudWatch metrics when this job runs. To see profiling data, you must enable this option. For more information about how to turn on and visualize metrics, see Job monitoring and debugging.

Job observability metrics

Turn on the creation of additional observability CloudWatch metrics when this job runs. For more information, see the section called “Monitoring with AWS Glue Observability metrics”.

Continuous logging

Turn on continuous logging to Amazon CloudWatch. If this option is not enabled, logs are available only after the job completes. For more information, see the section called “Logging for AWS Glue jobs”.

Spark UI

Turn on the use of Spark UI for monitoring this job. For more information, see Enabling the Apache Spark web UI for AWS Glue jobs.

Spark UI logs path

The path to write logs when Spark UI is enabled.

Spark UI logging and monitoring configuration

Choose one of the following options:

- *Standard*: write logs using the AWS Glue job run ID as the filename. Turn on Spark UI monitoring in the AWS Glue console.
- *Legacy*: write logs using 'spark-application-{timestamp}' as the filename. Do not turn on Spark UI monitoring.
- *Standard and legacy*: write logs to both the standard and legacy locations. Turn on Spark UI monitoring in the AWS Glue console.

Maximum concurrency

Sets the maximum number of concurrent runs that are allowed for this job. The default is 1. An error is returned when this threshold is reached. The maximum value you can specify is controlled by a service limit. For example, if a previous run of a job is still running when a new instance is started, you might want to return an error to prevent two instances of the same job from running concurrently.

Temporary path

Provide the location of a working directory in Amazon S3 where temporary intermediate results are written when AWS Glue runs the script. Confirm that there isn't a file with the same name as the temporary directory in the path. This directory is used when AWS Glue reads and writes to Amazon Redshift and by certain AWS Glue transforms.

> **Note**
>
> AWS Glue creates a temporary bucket for jobs if a bucket doesn't already exist in a region. This bucket might permit public access. You can either modify the bucket in Amazon S3 to set the public access block, or delete the bucket later after all jobs in that region have completed.

Delay notification threshold (minutes)

Sets the threshold (in minutes) before a delay notification is sent. You can set this threshold to send notifications when a `RUNNING`, `STARTING`, or `STOPPING` job run takes more than an expected number of minutes.

Security configuration

Choose a security configuration from the list. A security configuration specifies how the data at the Amazon S3 target is encrypted: no encryption, server-side encryption with AWS KMS-managed keys (SSE-KMS), or Amazon S3-managed encryption keys (SSE-S3).

Server-side encryption

If you select this option, when the ETL job writes to Amazon S3, the data is encrypted at rest using SSE-S3 encryption. Both your Amazon S3 data target and any data that is written to an Amazon S3 temporary directory is encrypted. This option is passed as a job parameter. For more information, see Protecting Data Using Server-Side Encryption with Amazon S3-Managed Encryption Keys (SSE-S3) in the *Amazon Simple Storage Service User Guide*.

> **Important**
>
> This option is ignored if a security configuration is specified.

Use Glue data catalog as the Hive metastore

Select to use the AWS Glue Data Catalog as the Hive metastore. The IAM role used for the job must have the `glue:CreateDatabase` permission. A database called “default” is created in the Data Catalog if it does not exist.

Connections

Choose a VPC configuration to access Amazon S3 data sources located in your virtual private cloud (VPC). You can create and manage Network connection in AWS Glue. For more information, see Connecting to data.

Libraries

Python library path, Dependent JARs path, and Referenced files path

Specify these options if your script requires them. You can define the comma-separated Amazon S3 paths for these options when you define the job. You can override these paths when you run the job. For more information, see Providing your own custom scripts.

Job parameters

A set of key-value pairs that are passed as named parameters to the script. These are default values that are used when the script is run, but you can override them in triggers or when you run the job. You must prefix the key name with `--`; for example: `--myKey`. You pass job parameters as a map when using the AWS Command Line Interface.

For examples, see Python parameters in Passing and accessing Python parameters in AWS Glue.

Tags

Tag your job with a **Tag key** and an optional **Tag value**. After tag keys are created, they are read-only. Use tags on some resources to help you organize and identify them. For more information, see AWS tags in AWS Glue.

Restrictions for jobs that access Lake Formation managed tables

Keep in mind the following notes and restrictions when creating jobs that read from or write to tables managed by AWS Lake Formation:

- The following features are not supported in jobs that access tables with cell-level filters:
  - Job bookmarks and bounded execution
  - Push-down predicates
  - Server-side catalog partition predicates
  - enableUpdateCatalog

Editing Spark scripts in the AWS Glue console

A script contains the code that extracts data from sources, transforms it, and loads it into targets. AWS Glue runs a script when it starts a job.

AWS Glue ETL scripts can be coded in Python or Scala. Python scripts use a language that is an extension of the PySpark Python dialect for extract, transform, and load (ETL) jobs. The script contains *extended constructs* to deal with ETL transformations. When you automatically generate the source code logic for your job, a script is created. You can edit this script, or you can provide your own script to process your ETL work.

For information about defining and editing scripts in AWS Glue, see AWS Glue programming guide.

Additional libraries or files

If your script requires additional libraries or files, you can specify them as follows:

Python library path

Comma-separated Amazon Simple Storage Service (Amazon S3) paths to Python libraries that are required by the script.

> **Note**
>
> Only pure Python libraries can be used. Libraries that rely on C extensions, such as the pandas Python Data Analysis Library, are not yet supported.

Dependent jars path

Comma-separated Amazon S3 paths to JAR files that are required by the script.

> **Note**
>
> Currently, only pure Java or Scala (2.11) libraries can be used.

Referenced files path

Comma-separated Amazon S3 paths to additional files (for example, configuration files) that are required by the script.

Jobs (legacy)

A script contains the code that performs extract, transform, and load (ETL) work. You can provide your own script, or AWS Glue can generate a script with guidance from you. For information about creating your own scripts, see Providing your own custom scripts.

You can edit a script in the AWS Glue console. When you edit a script, you can add sources, targets, and transforms.

To edit a script

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/. Then choose the **Jobs** tab.
2. Choose a job in the list, and then choose **Action**, **Edit script** to open the script editor.

   You can also access the script editor from the job details page. Choose the **Script** tab, and then choose **Edit script**.

Script editor

The AWS Glue script editor lets you insert, modify, and delete sources, targets, and transforms in your script. The script editor displays both the script and a diagram to help you visualize the flow of data.

To create a diagram for the script, choose **Generate diagram**. AWS Glue uses annotation lines in the script beginning with **##** to render the diagram. To correctly represent your script in the diagram, you must keep the parameters in the annotations and the parameters in the Apache Spark code in sync.

The script editor lets you add code templates wherever your cursor is positioned in the script. At the top of the editor, choose from the following options:

- To add a source table to the script, choose **Source**.
- To add a target table to the script, choose **Target**.
- To add a target location to the script, choose **Target location**.
- To add a transform to the script, choose **Transform**. For information about the functions that are called in your script, see Program AWS Glue ETL scripts in PySpark.
- To add a Spigot transform to the script, choose **Spigot**.

In the inserted code, modify the `parameters` in both the annotations and Apache Spark code. For example, if you add a **Spigot** transform, verify that the `path` is replaced in both the `@args` annotation line and the `output` code line.

The **Logs** tab shows the logs that are associated with your job as it runs. The most recent 1,000 lines are displayed.

The **Schema** tab shows the schema of the selected sources and targets, when available in the Data Catalog.

Tracking processed data using job bookmarks

AWS Glue tracks data that has already been processed during a previous run of an ETL job by persisting state information from the job run. This persisted state information is called a *job bookmark*. Job bookmarks help AWS Glue maintain state information and prevent the reprocessing of old data. With job bookmarks, you can process new data when rerunning on a scheduled interval. A job bookmark is composed of the states for various elements of jobs, such as sources, transformations, and targets. For example, your ETL job might read new partitions in an Amazon S3 file. AWS Glue tracks which partitions the job has processed successfully to prevent duplicate processing and duplicate data in the job's target data store.

Job bookmarks are implemented for JDBC data sources, the Relationalize transform, and some Amazon Simple Storage Service (Amazon S3) sources. The following table lists the Amazon S3 source formats that AWS Glue supports for job bookmarks.

AWS Glue version	Amazon S3 source formats
Version 0.9	JSON, CSV, Apache Avro, XML
Version 1.0 and later	JSON, CSV, Apache Avro, XML, Parquet, ORC

For information about AWS Glue versions, see Defining job properties for Spark jobs.

The job bookmarks feature has additional functionalities when accessed through AWS Glue scripts. When browsing your generated script, you may see transformation contexts, which are related to this feature. For more information, see the section called “Using job bookmarks”.

Topics

- Using job bookmarks in AWS Glue

- Operational details of the job bookmarks feature

Using job bookmarks in AWS Glue

The job bookmark option is passed as a parameter when the job is started. The following table describes the options for setting job bookmarks on the AWS Glue console.

Job bookmark	Description
Enable	Causes the job to update the state after a run to keep track of previously processed data. If your job has a source with job bookmark support, it will keep track of processed data, and when a job runs, it processes new data since the last checkpoint.
Disable	Job bookmarks are not used, and the job always processes the entire dataset. You are responsible for managing the output from previous job runs. This is the default.
Pause	Process incremental data since the last successful run or the data in the range identified by the following sub-options, without updating the state of last bookmark. You are responsible for managing the output from previous job runs. The two sub-options are: • **job-bookmark-from** <from-value> is the run ID which represents all the input that was processed until the last successful run before and including the specified run ID. The corresponding input is ignored. • **job-bookmark-to** <to-value> is the run ID which represents all the input that was processed until the last successful run before and including the specified run ID. The corresponding input excluding the input identified by the <from-value> is processed by the job. Any input later than this input is also excluded for processing. The job bookmark state is not updated when this option set is specified . The sub-options are optional, however when used both the sub-options needs to be provided.

For details about the parameters passed to a job on the command line, and specifically for job bookmarks, see Using job parameters in AWS Glue jobs.

For Amazon S3 input sources, AWS Glue job bookmarks check the last modified time of the objects to verify which objects need to be reprocessed. If your input source data has been modified since your last job run, the files are reprocessed when you run the job again.

For JDBC sources, the following rules apply:

- For each table, AWS Glue uses one or more columns as bookmark keys to determine new and processed data. The bookmark keys combine to form a single compound key.
- AWS Glue by default uses the primary key as the bookmark key, provided that it is sequentially increasing or decreasing (with no gaps).
- You can specify the columns to use as bookmark keys in your AWS Glue script. For more information about using Job bookmarks in AWS Glue scripts, see the section called “Using job bookmarks”.
- AWS Glue doesn't support using columns with case-sensitive names as job bookmark keys.

You can rewind your job bookmarks for your AWS Glue Spark ETL jobs to any previous job run. You can support data backfilling scenarios better by rewinding your job bookmarks to any previous job run, resulting in the subsequent job run reprocessing data only from the bookmarked job run.

If you intend to reprocess all the data using the same job, reset the job bookmark. To reset the job bookmark state, use the AWS Glue console, the ResetJobBookmark action (Python: reset_job_bookmark) API operation, or the AWS CLI. For example, enter the following command using the AWS CLI:

```
aws glue reset-job-bookmark --job-name my-job-name
```

When you rewind or reset a bookmark, AWS Glue does not clean the target files because there could be multiple targets and targets are not tracked with job bookmarks. Only source files are tracked with job bookmarks. You can create different output targets when rewinding and reprocessing the source files to avoid duplicate data in your output.

AWS Glue keeps track of job bookmarks by job. If you delete a job, the job bookmark is deleted.

In some cases, you might have enabled AWS Glue job bookmarks but your ETL job is reprocessing data that was already processed in an earlier run. For information about resolving common causes of this error, see Troubleshooting errors Spark errors.

Operational details of the job bookmarks feature

This section describes more of the operational details of using job bookmarks.

Job bookmarks store the states for a job. Each instance of the state is keyed by a job name and a version number. When a script invokes `job.init`, it retrieves its state and always gets the latest version. Within a state, there are multiple state elements, which are specific to each source, transformation, and sink instance in the script. These state elements are identified by a transformation context that is attached to the corresponding element (source, transformation, or sink) in the script. The state elements are saved atomically when `job.commit` is invoked from the user script. The script gets the job name and the control option for the job bookmarks from the arguments.

The state elements in the job bookmark are source, transformation, or sink-specific data. For example, suppose that you want to read incremental data from an Amazon S3 location that is being constantly written to by an upstream job or process. In this case, the script must determine what has been processed so far. The job bookmark implementation for the Amazon S3 source saves information so that when the job runs again, it can filter only the new objects using the saved information and recompute the state for the next run of the job. A timestamp is used to filter the new files.

In addition to the state elements, job bookmarks have a *run number*, an *attempt number*, and a *version number*. The run number tracks the run of the job, and the attempt number records the attempts for a job run. The job run number is a monotonically increasing number that is incremented for every successful run. The attempt number tracks the attempts for each run, and is only incremented when there is a run after a failed attempt. The version number increases monotonically and tracks the updates to a job bookmark.

In the AWS Glue service database, the bookmark states for all the transformations are stored together as key-value pairs:

```
{
  "job_name" : ...,
  "run_id": ...,
  "run_number": ..,
  "attempt_number": ...
  "states": {
    "transformation_ctx1" : {
      bookmark_state1
    },
    "transformation_ctx2" : {
      bookmark_state2
    }
  }
}
```

Best practices

The following are best practices for using job bookmarks.

- *Do not change the data source property with the bookmark enabled*. For example, there is a datasource0 pointing to an Amazon S3 input path A, and the job has been reading from a source which has been running for several rounds with the bookmark enabled. If you change the input path of datasource0 to Amazon S3 path B without changing the `transformation_ctx`, the AWS Glue job will use the old bookmark state stored. That will result in missing or skipping files in the input path B as AWS Glue would assume that those files had been processed in previous runs.
- *Use a catalog table with bookmarks for better partition management*. Bookmarks works both for data sources from the Data Catalog or from options. However, it's difficult to remove/add new partitions with the from options approach. Using a catalog table with crawlers can provide better automation to track the newly added partitions and give you the flexibility to select particular partitions with a pushdown predicate.
- *Use the AWS Glue Amazon S3 file lister for large datasets*. A bookmark will list all files under each input partition and do the filering, so if there are too many files under a single partition the bookmark can run into driver OOM. Use the AWS Glue Amazon S3 file lister to avoid listing all files in memory at once.

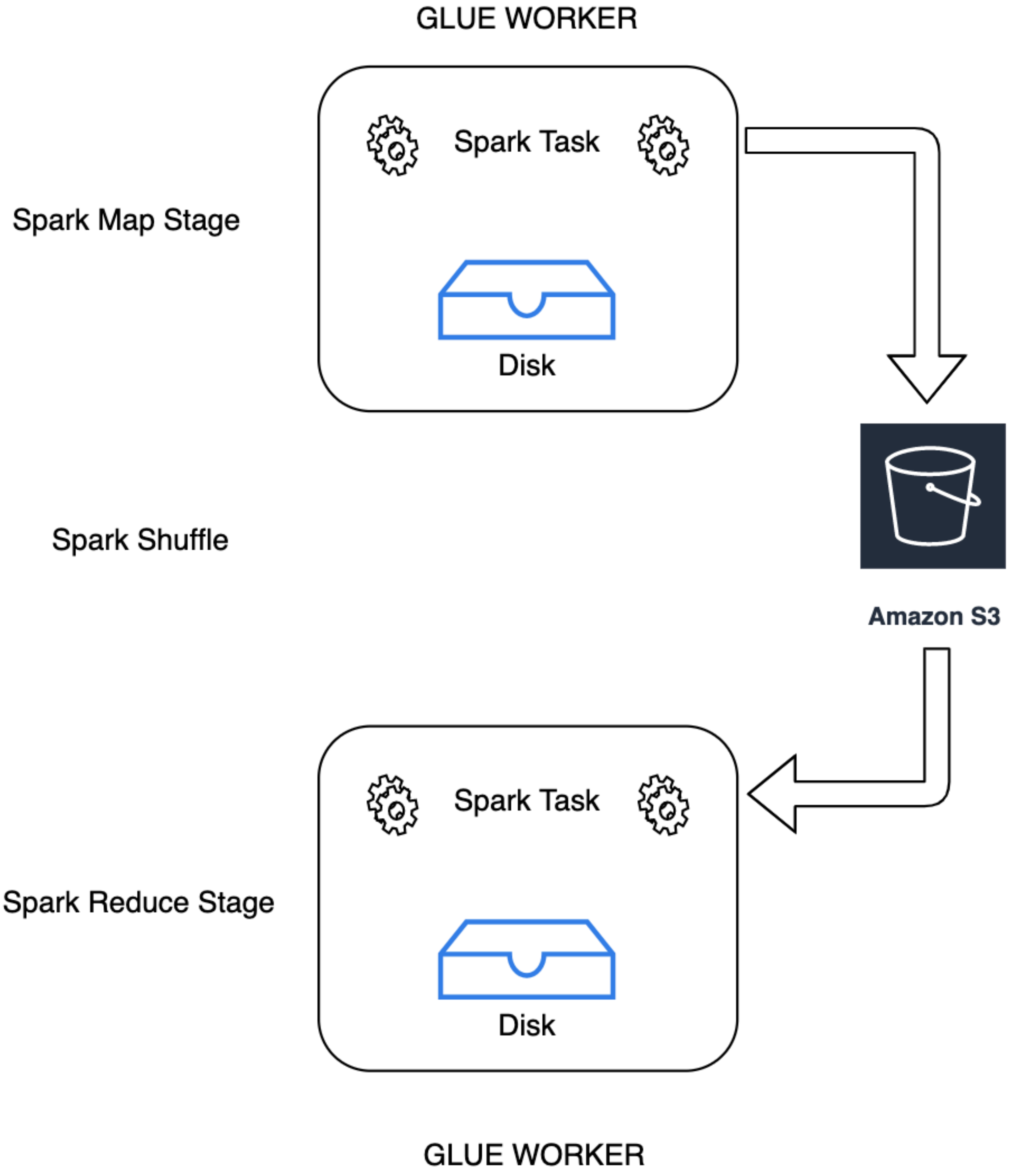
Storing Spark shuffle data

Shuffling is an important step in a Spark job whenever data is rearranged between partitions. This is required because wide transformations such as `join`, `groupByKey`, `reduceByKey`, and `repartition` require information from other partitions to complete processing. Spark gathers the required data from each partition and combines it into a new partition. During a shuffle, data is written to disk and transferred across the network. As a result, the shuffle operation is bound to local disk capacity. Spark throws a `No space left on device` or `MetadataFetchFailedException` error when there is not enough disk space left on the executor and there is no recovery.

> **Note**
>
> AWS Glue Spark shuffle plugin with Amazon S3 is only supported for AWS Glue ETL jobs.

Solution

With AWS Glue, you can now use Amazon S3 to store Spark shuffle data. Amazon S3 is an object storage service that offers industry-leading scalability, data availability, security, and performance. This solution disaggregates compute and storage for your Spark jobs, and gives complete elasticity and low-cost shuffle storage, allowing you to run your most shuffle-intensive workloads reliably.


GLUE WORKER
Spark Task
Disk
Spark Map Stage
Spark Shuffle
Amazon S3
Spark Task
Disk
Spark Reduce Stage
GLUE WORKER

We are introducing a new Cloud Shuffle Storage Plugin for Apache Spark to use Amazon S3. You can turn on Amazon S3 shuffling to run your AWS Glue jobs reliably without failures if they are known to be bound by the local disk capacity for large shuffle operations. In some cases, shuffling to Amazon S3 is marginally slower than local disk (or EBS) if you have a large number of small partitions or shuffle files written out to Amazon S3.

Prerequisites for using Cloud Shuffle Storage Plugin

In order to use the Cloud Shuffle Storage Plugin with AWS Glue ETL jobs, you need the following:

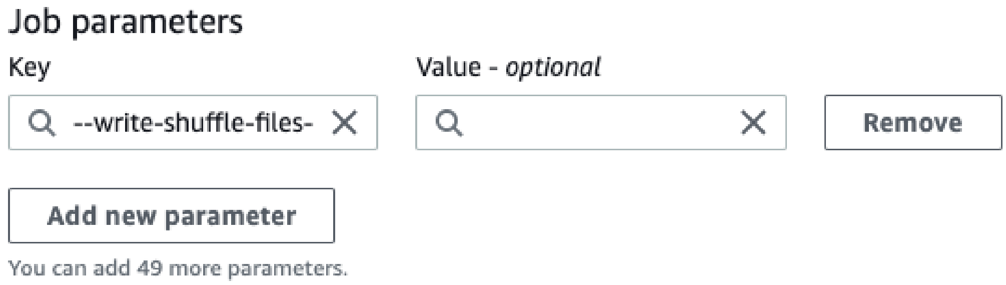
- An Amazon S3 bucket located in the same region as your job run, for storing the intermediate shuffle and spilled data. The Amazon S3 prefix of shuffle storage can be specified with `--conf spark.shuffle.glue.s3ShuffleBucket=s3://shuffle-bucket/prefix/`, as in the following example:

```
--conf spark.shuffle.glue.s3ShuffleBucket=s3://glue-shuffle-123456789-us-east-1/glue-shuffle-data/
```

- Set the Amazon S3 storage lifecycle policies on the *prefix* (such as `glue-shuffle-data`) as the shuffle manager does not clean the files after the job is done. The intermediate shuffle and spilled data should be deleted after a job is finished. Users can set a short lifecycle policies on the prefix. Instructions for setting up an Amazon S3 lifecycle policy are available at Setting lifecycle configuration on a bucket in the Amazon Simple Storage Service User Guide.

Using AWS Glue Spark shuffle manager from the AWS console

To set up the AWS Glue Spark shuffle manager using the AWS Glue console or AWS Glue Studio when configuring a job: choose the **--write-shuffle-files-to-s3** job parameter to turn on Amazon S3 shuffling for the job.


Job parameters

Key: --write-shuffle-files-

Value - *optional*:

Using AWS Glue Spark shuffle plugin

The following job parameters turn on and tune the AWS Glue shuffle manager. These parameters are flags, so any values provided are not considered.

- `--write-shuffle-files-to-s3` — The main flag, which enables the AWS Glue Spark shuffle manager to use Amazon S3 buckets for writing and reading shuffle data. When the flag is not specified, the shuffle manager is not used.
- `--write-shuffle-spills-to-s3` — (Supported only on AWS Glue version 2.0). An optional flag that allows you to offload spill files to Amazon S3 buckets, which provides additional resiliency to your Spark job. This is only required for large workloads that spill a lot of data to disk. When the flag is not specified, no intermediate spill files are written.
- `--conf spark.shuffle.glue.s3ShuffleBucket=s3://<shuffle-bucket>` — Another optional flag that specifies the Amazon S3 bucket where you write the shuffle files. By default, `--TempDir`/shuffle-data. AWS Glue 3.0+ supports writing shuffle files to multiple buckets by specifying buckets with comma delimiter, as in `--conf spark.shuffle.glue.s3ShuffleBucket=s3://shuffle-bucket-1/prefix,s3://shuffle-bucket-2/prefix/`. Using multiple buckets improves performance.

You need to provide security configuration settings to enable encryption at-rest for the shuffle data. For more information about security configurations, see the section called “Setting up encryption”. AWS Glue supports all other shuffle related configurations provided by Spark.

Software binaries for the Cloud Shuffle Storage plugin

You can also download the software binaries of the Cloud Shuffle Storage Plugin for Apache Spark under the Apache 2.0 license and run it on any Spark environment. The new plugin comes with out-of-the box support for Amazon S3, and can also be easily configured to use other forms of cloud storage such as Google Cloud Storage and Microsoft Azure Blob Storage. For more information, see Cloud Shuffle Storage Plugin for Apache Spark.

Notes and limitations

The following are notes or limitations for the AWS Glue shuffle manager:

- AWS Glue shuffle manager doesn't automatically delete the (temporary) shuffle data files stored in your Amazon S3 bucket after a job is completed. To ensure data protection, follow the instructions in Prerequisites for using Cloud Shuffle Storage Plugin before enabling the Cloud Shuffle Storage Plugin.
- You can use this feature if your data is skewed.

Cloud Shuffle Storage Plugin for Apache Spark

The Cloud Shuffle Storage Plugin is an Apache Spark plugin compatible with the `ShuffleDataIO` API which allows storing shuffle data on cloud storage systems (such as Amazon S3). It helps you to supplement or replace local disk storage capacity for large shuffle operations, commonly triggered by transformations such as `join`, `reduceByKey`, `groupByKey` and `repartition` in your Spark applications, thereby reducing common failures or price/performance dislocation of your serverless data analytics jobs and pipelines.

AWS Glue

AWS Glue versions 3.0 and 4.0 comes with the plugin pre-installed and ready to enable shuffling to Amazon S3 without any extra steps. For more information, see AWS Glue Spark shuffle plugin with Amazon S3 to enable the feature for your Spark applications.

Other Spark environments

The plugin requires the following Spark configurations to be set on other Spark environments:

- `--conf spark.shuffle.sort.io.plugin.class=com.amazonaws.spark.shuffle.io.cloud.Choppe` This informs Spark to use this plugin for Shuffle IO.
- `--conf spark.shuffle.storage.path=s3://bucket-name/shuffle-file-dir`: The path where your shuffle files will be stored.

> **Note**
>
> The plugin overwrites one Spark core class. As a result, the plugin jar needs to be loaded before Spark jars. You can do this using `userClassPathFirst` in on-prem YARN environments if the plugin is used outside AWS Glue.

Bundling the plugin with your Spark applications

You can bundle the plugin with your Spark applications and Spark distributions (versions 3.1 and above) by adding the plugin dependency in your Maven `pom.xml` while developing your Spark applications locally. For more information on the plugin and Spark versions, see Plugin versions.

```
<repositories>
```

```
   ...
    <repository>
        <id>aws-glue-etl-artifacts</id>
        <url>https://aws-glue-etl-artifacts.s3.amazonaws.com/release/ </url>
    </repository>
</repositories>
...
<dependency>
    <groupId>com.amazonaws</groupId>
    <artifactId>chopper-plugin</artifactId>
    <version>3.1-amzn-LATEST</version>
</dependency>
```

You can alternatively download the binaries from AWS Glue Maven artifacts directly and include them in your Spark application as follows.

```
#!/bin/bash
sudo wget -v https://aws-glue-etl-artifacts.s3.amazonaws.com/release/com/amazonaws/
chopper-plugin/3.1-amzn-LATEST/chopper-plugin-3.1-amzn-LATEST.jar -P /usr/lib/spark/
jars/
```

Example spark-submit

```
spark-submit --deploy-mode cluster \
--conf spark.shuffle.storage.s3.path=s3://<ShuffleBucket>/<shuffle-dir> \
--conf spark.driver.extraClassPath=<Path to plugin jar> \
--conf spark.executor.extraClassPath=<Path to plugin jar> \
--class <your test class name> s3://<ShuffleBucket>/<Your application jar> \
```

Optional configurations

These are optional configuration values that control Amazon S3 shuffle behavior.

- `spark.shuffle.storage.s3.enableServerSideEncryption`: Enable/disable S3 SSE for shuffle and spill files. Default value is `true`.
- `spark.shuffle.storage.s3.serverSideEncryption.algorithm`: The SSE algorithm to be used. Default value is `AES256`.
- `spark.shuffle.storage.s3.serverSideEncryption.kms.key`: The KMS key ARN when SSE aws:kms is enabled.

Along with these configurations, you may need to set configurations such as `spark.hadoop.fs.s3.enableServerSideEncryption` and **other environment-specific configurations** to ensure appropriate encryption is applied for your use case.

Plugin versions

This plugin is supported for the Spark versions associated with each AWS Glue version. The following table shows the AWS Glue version, Spark version and associated plugin version with Amazon S3 location for the plugin's software binary.

AWS Glue version	Spark version	Plugin version	Amazon S3 location
3.0	3.1	3.1-amzn-LATEST	s3://aws-glue-etl-artifacts/release/com/amazonaws/chopper-plugin/3.1-amzn-0/chopper-plugin-3.1-amzn-LATEST.jar
4.0	3.3	3.3-amzn-LATEST	s3://aws-glue-etl-artifacts/release/com/amazonaws/chopper-plugin/3.3-amzn-0/chopper-plugin-3.3-amzn-LATEST.jar

License

The software binary for this plugin is licensed under the Apache-2.0 License.

Monitoring AWS Glue Spark jobs

Topics

- Spark Metrics available in AWS Glue Studio
- Monitoring jobs using the Apache Spark web UI

- Monitoring with AWS Glue job run insights
- Monitoring with Amazon CloudWatch
- Job monitoring and debugging

Spark Metrics available in AWS Glue Studio

The **Metrics** tab shows metrics collected when a job runs and profiling is turned on. The following graphs are shown in Spark jobs:

- ETL Data Movement
- Memory Profile: Driver and Executors

Choose **View additional metrics** to show the following graphs:

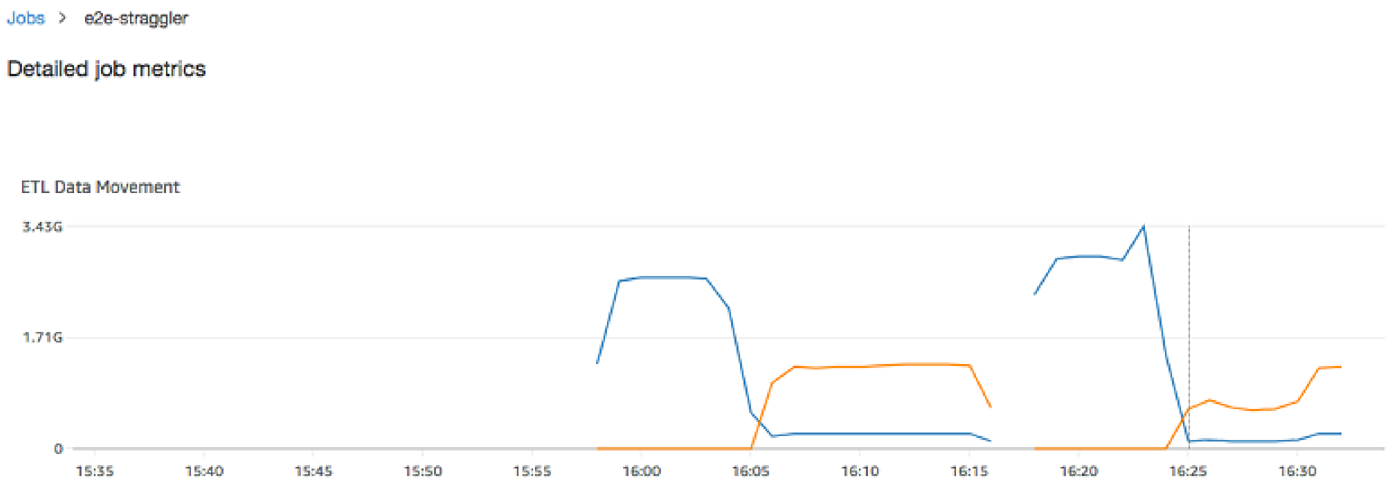
- ETL Data Movement
- Memory Profile: Driver and Executors
- Data Shuffle Across Executors
- CPU Load: Driver and Executors
- Job Execution: Active Executors, Completed Stages & Maximum Needed Executors

Data for these graphs is pushed to CloudWatch metrics if the job is configured to collect metrics. For more information about how to turn on metrics and interpret the graphs, see Job monitoring and debugging.

Example ETL data movement graph

The ETL Data Movement graph shows the following metrics:

- The number of bytes read from Amazon S3 by all executors —`glue.ALL.s3.filesystem.read_bytes`
- The number of bytes written to Amazon S3 by all executors —`glue.ALL.s3.filesystem.write_bytes`

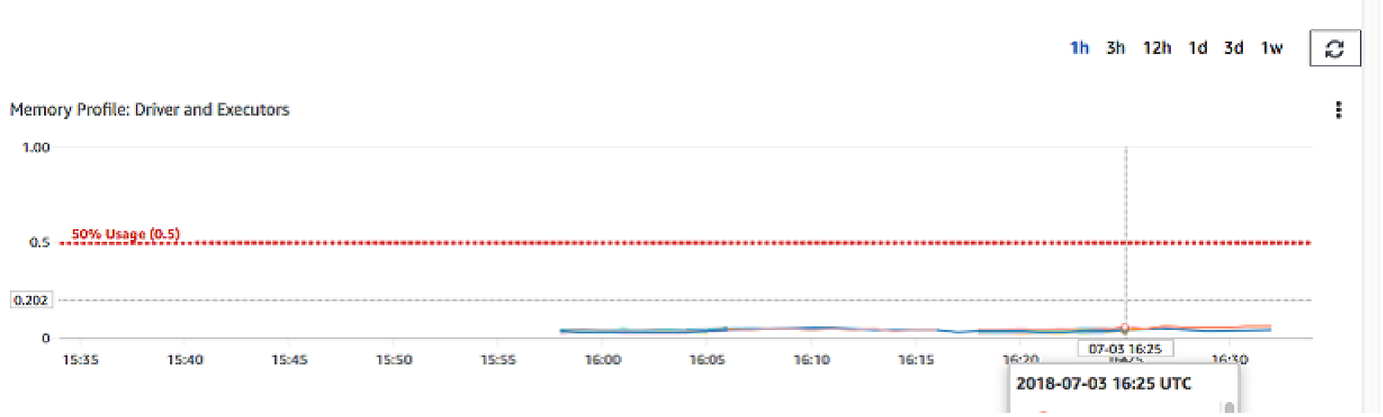

Jobs > e2e-straggler

Detailed job metrics

Example Memory profile graph

The Memory Profile graph shows the following metrics:

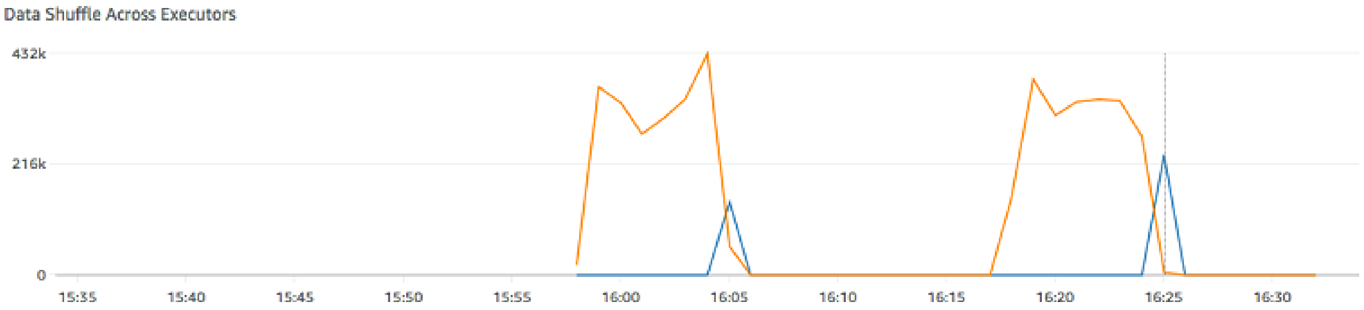
- The fraction of memory used by the JVM heap for this driver (scale: 0–1) by the driver, an executor identified by *executorId*, or all executors—
  - `glue.driver.jvm.heap.usage`
  - `glue.`*`executorId`*`.jvm.heap.usage`
  - `glue.ALL.jvm.heap.usage`



Example Data shuffle across executors graph

The Data Shuffle Across Executors graph shows the following metrics:

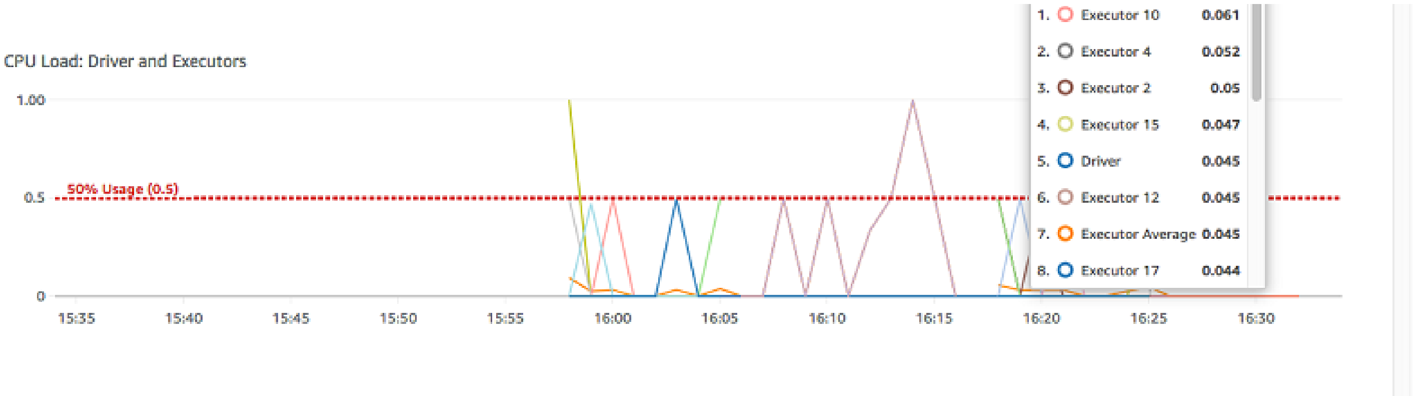
- The number of bytes read by all executors to shuffle data between them —`glue.driver.aggregate.shuffleLocalBytesRead`
- The number of bytes written by all executors to shuffle data between them —`glue.driver.aggregate.shuffleBytesWritten`



Example CPU load graph

The CPU Load graph shows the following metrics:

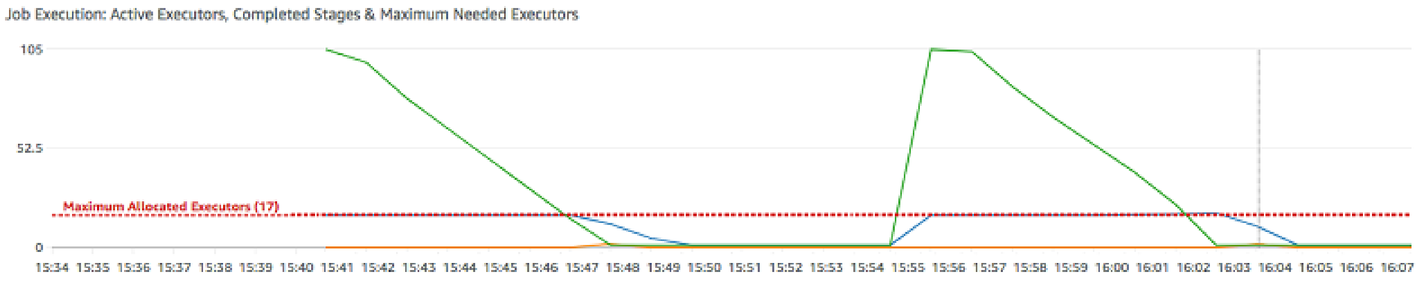
- The fraction of CPU system load used (scale: 0–1) by the driver, an executor identified by *executorId*, or all executors—
  - `glue.driver.system.cpuSystemLoad`
  - `glue.`*`executorId`*`.system.cpuSystemLoad`
  - `glue.ALL.system.cpuSystemLoad`



Example Job execution graph

The Job Execution graph shows the following metrics:

- The number of actively running executors —`glue.driver.ExecutorAllocationManager.executors.numberAllExecutors`
- The number of completed stages—`glue.aggregate.numCompletedStages`
- The number of maximum needed executors —`glue.driver.ExecutorAllocationManager.executors.numberMaxNeededExecutors`



Monitoring jobs using the Apache Spark web UI

You can use the Apache Spark web UI to monitor and debug AWS Glue ETL jobs running on the AWS Glue job system, and also Spark applications running on AWS Glue development endpoints. The Spark UI enables you to check the following for each job:

- The event timeline of each Spark stage
- A directed acyclic graph (DAG) of the job
- Physical and logical plans for SparkSQL queries
- The underlying Spark environmental variables for each job

For more information about using the Spark Web UI, see Web UI in the Spark documentation. For guidance on how to interpret Spark UI results to improve the performance of your job, see Best practices for performance tuning AWS Glue for Apache Spark jobs in AWS Prescriptive Guidance.

You can see the Spark UI in the AWS Glue console. This is available when an AWS Glue job runs on AWS Glue 3.0 or later versions with logs generated in the Standard (rather than legacy) format, which is the default for newer jobs. If you have log files greater than 0.5 GB, you can enable rolling log support for job runs on AWS Glue 4.0 or later versions to simplify log archiving, analysis, and troubleshooting.

You can enable the Spark UI by using the AWS Glue console or the AWS Command Line Interface (AWS CLI). When you enable the Spark UI, AWS Glue ETL jobs and Spark applications on AWS Glue development endpoints can back up Spark event logs to a location that you specify in Amazon Simple Storage Service (Amazon S3). You can use the backed up event logs in Amazon S3 with the Spark UI, both in real time as the job is operating and after the job is complete. While the logs remain in Amazon S3, the Spark UI in the AWS Glue console can view them.

Permissions

In order to use the Spark UI in the AWS Glue console, you can use `UseGlueStudio` or add all the individual service APIs. All APIs are needed to use the Spark UI completely, however users can access SparkUI features by adding its service APIs in their IAM permission for fine-grained access.

`RequestLogParsing` is the most critical as it performs the parsing of logs. The remaining APIs are for reading the respective parsed data. For example, `GetStages` provides access to the data about all stages of a Spark job.

The list of Spark UI service APIs mapped to `UseGlueStudio` are below in the sample policy. The policy below provides access to use only Spark UI features. To add more permissions like Amazon S3 and IAM see [Creating Custom IAM Policies for AWS Glue Studio.]

The list of Spark UI service APIs mapped to `UseGlueStudio` is below in the sample policy. When using a Spark UI service API, use the following namespace: `glue:<ServiceAPI>`.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "AllowGlueStudioSparkUI",
      "Effect": "Allow",
      "Action": [
        "glue:RequestLogParsing",
        "glue:GetLogParsingStatus",
        "glue:GetEnvironment",
        "glue:GetJobs",
        "glue:GetJob",
        "glue:GetStage",
        "glue:GetStages",
        "glue:GetStageFiles",
        "glue:BatchGetStageFiles",
        "glue:GetStageAttempt",
          "glue:GetStageAttemptTaskList",
          "glue:GetStageAttemptTaskSummary",
          "glue:GetExecutors",
          "glue:GetExecutorsThreads",
          "glue:GetStorage",
          "glue:GetStorageUnit",
          "glue:GetQueries",
          "glue:GetQuery"
        ],
        "Resource": [
          "*"
        ]
      }
    ]
  }
```

Limitations

- Spark UI in the AWS Glue console is not available for job runs that occurred before 20 Nov 2023 because they are in the legacy log format.
- Spark UI in the AWS Glue console supports rolling logs for AWS Glue 4.0, such as those generated by default in streaming jobs. The maximum sum of all generated rolled log event files is 2 GB. For AWS Glue jobs without rolled log support, the maximum log event file size supported for SparkUI is 0.5 GB.
- Serverless Spark UI is not available for Spark event logs stored in an Amazon S3 bucket that can only be accessed by your VPC.

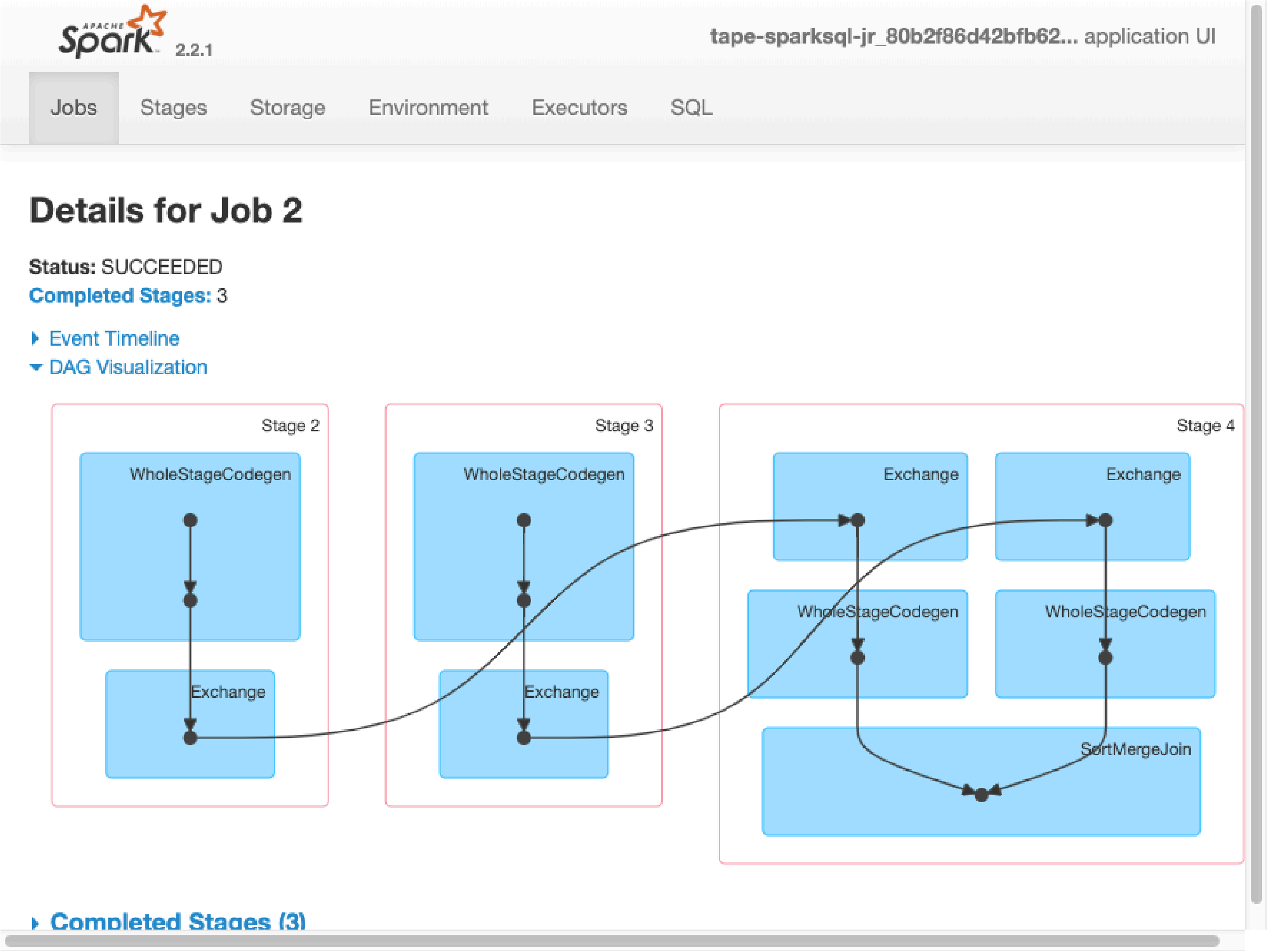
Example: Apache Spark web UI

This example shows you how to use the Spark UI to understand your job performance. Screen shots show the Spark web UI as provided by a self-managed Spark history server. Spark UI in the AWS Glue console provides similar views. For more information about using the Spark Web UI, see Web UI in the Spark documentation.

The following is an example of a Spark application that reads from two data sources, performs a join transform, and writes it out to Amazon S3 in Parquet format.

```
import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job
from pyspark.sql.functions import count, when, expr, col, sum, isnull
from pyspark.sql.functions import countDistinct
from awsglue.dynamicframe import DynamicFrame

args = getResolvedOptions(sys.argv, ['JOB_NAME'])

sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session

job = Job(glueContext)
job.init(args['JOB_NAME'])

df_persons = spark.read.json("s3://awsglue-datasets/examples/us-legislators/all/
persons.json")
df_memberships = spark.read.json("s3://awsglue-datasets/examples/us-legislators/all/
memberships.json")

df_joined = df_persons.join(df_memberships, df_persons.id == df_memberships.person_id,
 'fullouter')
df_joined.write.parquet("s3://aws-glue-demo-sparkui/output/")

job.commit()
```

The following DAG visualization shows the different stages in this Spark job.



Spark 2.2.1
Jobs
Stages
Storage
Environment
Executors
SQL
Stage 2
Stage 3
Stage 4
WholeStageCodegen
Exchange
SortMergeJoin

tape-sparksql-jr_80b2f86d42bfb62... application UI

Details for Job 2

Status: SUCCEEDED

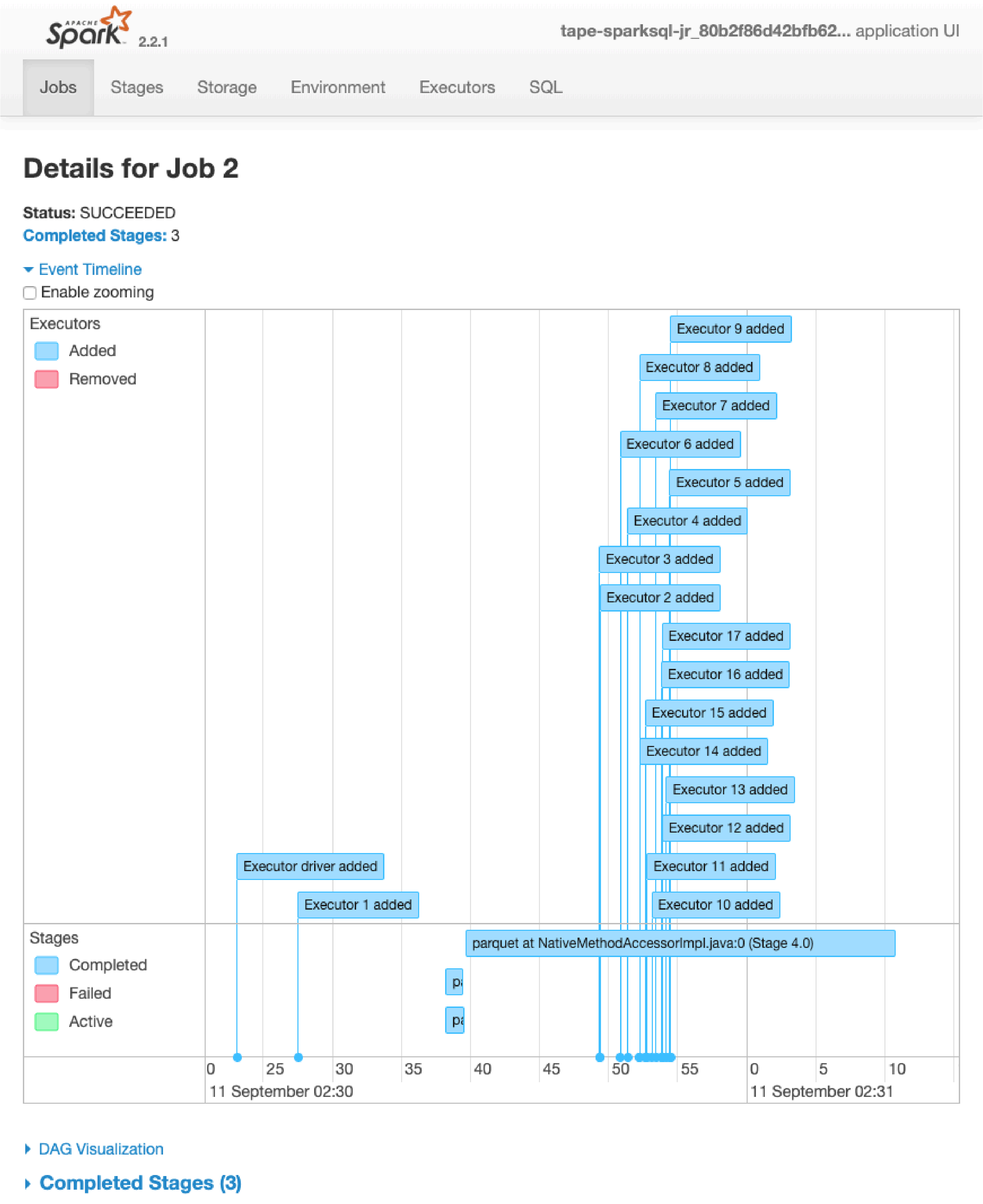
Completed Stages: 3

Event Timeline

DAG Visualization

Completed Stages (3)

The following event timeline for a job shows the start, execution, and termination of different Spark executors.



APACHE Spark 2.2.1
tape-sparksql-jr_80b2f86d42bfb62... application UI
Jobs
Stages
Storage
Environment
Executors
SQL

Details for Job 2

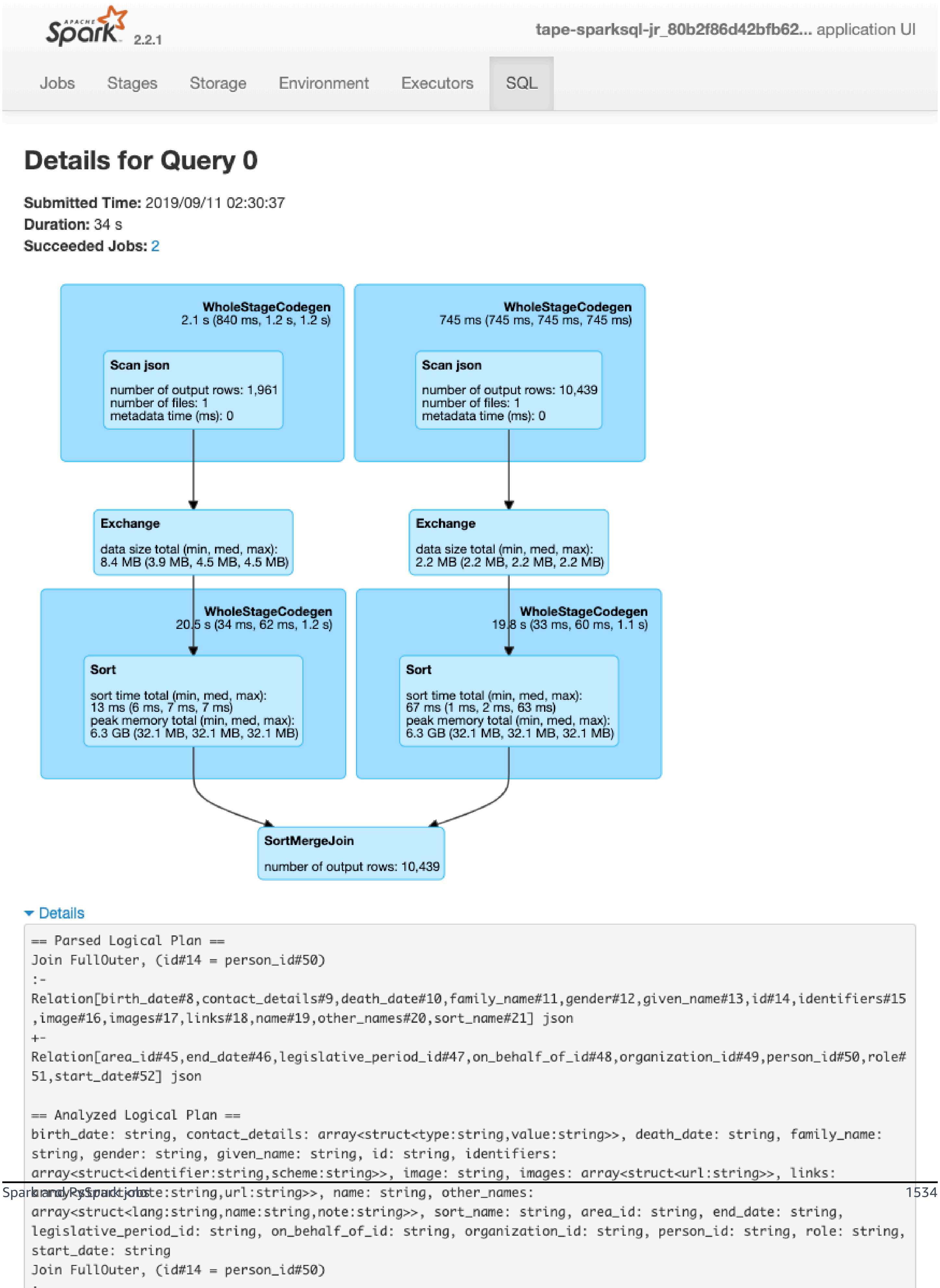
Status: SUCCEEDED

Completed Stages: 3

Event Timeline
☐ Enable zooming
Executors
Added
Removed
Executor 9 added
Executor 8 added
Executor 7 added
Executor 6 added
Executor 5 added
Executor 4 added
Executor 3 added
Executor 2 added
Executor 17 added
Executor 16 added
Executor 15 added
Executor 14 added
Executor 13 added
Executor 12 added
Executor driver added
Executor 11 added
Executor 1 added
Executor 10 added
Stages
Completed
Failed
Active
parquet at NativeMethodAccessorImpl.java:0 (Stage 4.0)
pa
pa
0
25
30
35
40
45
50
55
0
5
10
11 September 02:30
11 September 02:31
DAG Visualization
Completed Stages (3)

The following screen shows the details of the SparkSQL query plans:

- Parsed logical plan
- Analyzed logical plan
- Optimized logical plan
- Physical plan for execution



Details for Query 0

Submitted Time: 2019/09/11 02:30:37
Duration: 34 s
Succeeded Jobs: 2

WholeStageCodegen
2.1 s (840 ms, 1.2 s, 1.2 s)

Scan json
number of output rows: 1,961
number of files: 1
metadata time (ms): 0

WholeStageCodegen
745 ms (745 ms, 745 ms, 745 ms)

Scan json
number of output rows: 10,439
number of files: 1
metadata time (ms): 0

Exchange
data size total (min, med, max):
8.4 MB (3.9 MB, 4.5 MB, 4.5 MB)

Exchange
data size total (min, med, max):
2.2 MB (2.2 MB, 2.2 MB, 2.2 MB)

WholeStageCodegen
20.5 s (34 ms, 62 ms, 1.2 s)

Sort
sort time total (min, med, max):
13 ms (6 ms, 7 ms, 7 ms)
peak memory total (min, med, max):
6.3 GB (32.1 MB, 32.1 MB, 32.1 MB)

WholeStageCodegen
19.8 s (33 ms, 60 ms, 1.1 s)

Sort
sort time total (min, med, max):
67 ms (1 ms, 2 ms, 63 ms)
peak memory total (min, med, max):
6.3 GB (32.1 MB, 32.1 MB, 32.1 MB)

SortMergeJoin
number of output rows: 10,439

▾ Details

```
== Parsed Logical Plan ==
Join FullOuter, (id#14 = person_id#50)
:-
Relation[birth_date#8,contact_details#9,death_date#10,family_name#11,gender#12,given_name#13,id#14,identifiers#15
,image#16,images#17,links#18,name#19,other_names#20,sort_name#21] json
+-
Relation[area_id#45,end_date#46,legislative_period_id#47,on_behalf_of_id#48,organization_id#49,person_id#50,role#
51,start_date#52] json

== Analyzed Logical Plan ==
birth_date: string, contact_details: array<struct<type:string,value:string>>, death_date: string, family_name:
string, gender: string, given_name: string, id: string, identifiers:
array<struct<identifier:string,scheme:string>>, image: string, images: array<struct<url:string>>, links:
array<struct<note:string,url:string>>, name: string, other_names:
array<struct<lang:string,name:string,note:string>>, sort_name: string, area_id: string, end_date: string,
legislative_period_id: string, on_behalf_of_id: string, organization_id: string, person_id: string, role: string,
start_date: string
Join FullOuter, (id#14 = person_id#50)
```

Topics

- Enabling the Apache Spark web UI for AWS Glue jobs
- Launching the Spark history server

Enabling the Apache Spark web UI for AWS Glue jobs

You can use the Apache Spark web UI to monitor and debug AWS Glue ETL jobs running on the AWS Glue job system. You can configure the Spark UI using the AWS Glue console or the AWS Command Line Interface (AWS CLI).

Every 30 seconds, AWS Glue backs up the Spark event logs to the Amazon S3 path that you specify.

Topics

- Configuring the Spark UI (console)
- Configuring the Spark UI (AWS CLI)
- Configuring the Spark UI for sessions using Notebooks
- Enable rolling logs

Configuring the Spark UI (console)

Follow these steps to configure the Spark UI by using the AWS Management Console. When creating an AWS Glue job, Spark UI is enabled by default.

To turn on the Spark UI when you create or edit a job

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, choose **Jobs**.
3. Choose **Add job**, or select an existing one.
4. In **Job details**, open the **Advanced properties**.
5. Under the **Spark UI** tab, choose **Write Spark UI logs to Amazon S3**.
6. Specify an Amazon S3 path for storing the Spark event logs for the job. Note that if you use a security configuration in the job, the encryption also applies to the Spark UI log file. For more information, see Encrypting data written by AWS Glue.
7. Under **Spark UI logging and monitoring configuration**:

  - Select **Standard** if you are generating logs to view in the AWS Glue console.
  - Select **Legacy** if you are generating logs to view on a Spark history server.
  - You can also choose to generate both.

Configuring the Spark UI (AWS CLI)

To generate logs for viewing with Spark UI, in the AWS Glue console, use the AWS CLI to pass the following job parameters to AWS Glue jobs. For more information, see the section called “Job parameters”.

```
'--enable-spark-ui': 'true',
'--spark-event-logs-path': 's3://s3-event-log-path'
```

To distribute logs to their legacy locations, set the `--enable-spark-ui-legacy-path` parameter to `"true"`. If you do not want to generate logs in both formats, remove the `--enable-spark-ui` parameter.

Configuring the Spark UI for sessions using Notebooks

> **Warning**
>
> AWS Glue interactive sessions do not currently support Spark UI in the console. Configure a Spark history server.

If you use AWS Glue notebooks, set up SparkUI config before starting the session. To do this, use the `%%configure` cell magic:

```
%%configure { “--enable-spark-ui”: “true”, “--spark-event-logs-path”: “s3://path” }
```

Enable rolling logs

Enabling SparkUI and rolling log event files for AWS Glue jobs provides several benefits:

- Rolling Log Event Files – With rolling log event files enabled, AWS Glue generates separate log files for each step of the job execution, making it easier to identify and troubleshoot issues specific to a particular stage or transformation.

- Better Log Management – Rolling log event files help in managing log files more efficiently. Instead of having a single, potentially large log file, the logs are split into smaller, more manageable files based on the job execution stages. This can simplify log archiving, analysis, and troubleshooting.
- Improved Fault Tolerance – If a AWS Glue job fails or is interrupted, the rolling log event files can provide valuable information about the last successful stage, making it easier to resume the job from that point rather than starting from scratch.
- Cost Optimization – By enabling rolling log event files, you can save on storage costs associated with log files. Instead of storing a single, potentially large log file, you store smaller, more manageable log files, which can be more cost-effective, especially for long-running or complex jobs.

In a new environment, users can explicitly enable rolling logs through:

```
'–conf': 'spark.eventLog.rolling.enabled=true'
```

or

```
'–conf': 'spark.eventLog.rolling.enabled=true –conf
spark.eventLog.rolling.maxFileSize=128m'
```

When rolling logs are activated, `spark.eventLog.rolling.maxFileSize` specifies the maximum size of the event log file before it rolls over. The default value of this optional parameter if not specified is 128 MB. Minimum is 10 MB.

The maximum sum of all generated rolled log event files is 2 GB. For AWS Glue jobs without rolling log support, the maximum log event file size supported for SparkUI is 0.5 GB.

You can turn off rolling logs for a streaming job by passing an additional configuration. Note that very large log files may be costly to maintain.

To turn off rolling logs, provide the following configuration:

```
'--spark-ui-event-logs-path': 'true',
'--conf': 'spark.eventLog.rolling.enabled=false'
```

Launching the Spark history server

You can use a Spark history server to visualize Spark logs on your own infrastructure. You can see the same visualizations in the AWS Glue console for AWS Glue job runs on AWS Glue 4.0 or later versions with logs generated in the Standard (rather than legacy) format. For more information, see the section called “Monitoring with the Spark UI”.

You can launch the Spark history server using a AWS CloudFormation template that hosts the server on an EC2 instance, or launch locally using Docker.

Topics

- Launching the Spark history server and viewing the Spark UI using AWS CloudFormation
- Launching the Spark history server and viewing the Spark UI using Docker

Launching the Spark history server and viewing the Spark UI using AWS CloudFormation

You can use an AWS CloudFormation template to start the Apache Spark history server and view the Spark web UI. These templates are samples that you should modify to meet your requirements.

To start the Spark history server and view the Spark UI using AWS CloudFormation

1. Choose one of the **Launch Stack** buttons in the following table. This launches the stack on the AWS CloudFormation console.

Region	Launch
US East (Ohio)	Launch Stack
US East (N. Virginia)	Launch Stack
US West (N. California)	Launch Stack
US West (Oregon)	Launch Stack
Africa (Cape Town)	Launch Stack

Region	Launch
Asia Pacific (Hong Kong)	Launch Stack
Asia Pacific (Mumbai)	Launch Stack
Asia Pacific (Osaka)	Launch Stack
Asia Pacific (Seoul)	Launch Stack
Asia Pacific (Singapore)	Launch Stack
Asia Pacific (Sydney)	Launch Stack
Asia Pacific (Tokyo)	Launch Stack
Canada (Central)	Launch Stack
Europe (Frankfurt)	Launch Stack
Europe (Ireland)	Launch Stack
Europe (London)	Launch Stack
Europe (Milan)	Launch Stack
Europe (Paris)	Launch Stack
Europe (Stockholm)	Launch Stack
Middle East (Bahrain)	Launch Stack

Region	Launch
South America (São Paulo)	Launch Stack

2. On the **Specify template** page, choose **Next**.
3. On the **Specify stack details** page, enter the **Stack name**. Enter additional information under **Parameters**.

   a. **Spark UI configuration**

   Provide the following information:

   - **IP address range** — The IP address range that can be used to view the Spark UI. If you want to restrict access from a specific IP address range, you should use a custom value.
   - **History server port** — The port for the Spark UI. You can use the default value.
   - **Event log directory** — Choose the location where Spark event logs are stored from the AWS Glue job or development endpoints. You must use **`s3a://`** for the event logs path scheme.
   - **Spark package location** — You can use the default value.
   - **Keystore path** — SSL/TLS keystore path for HTTPS. If you want to use a custom keystore file, you can specify the S3 path `s3://path_to_your_keystore_file` here. If you leave this parameter empty, a self-signed certificate based keystore is generated and used.
   - **Keystore password** — Enter a SSL/TLS keystore password for HTTPS.

   b. **EC2 instance configuration**

   Provide the following information:

   - **Instance type** — The type of Amazon EC2 instance that hosts the Spark history server. Because this template launches Amazon EC2 instance in your account, Amazon EC2 cost will be charged in your account separately.
   - **Latest AMI ID** — The AMI ID of Amazon Linux 2 for the Spark history server instance. You can use the default value.
   - **VPC ID** — The virtual private cloud (VPC) ID for the Spark history server instance. You can use any of the VPCs available in your account Using a default VPC with a default

    Network ACL is not recommended. For more information, see Default VPC and Default Subnets and Creating a VPC in the *Amazon VPC User Guide*.

    - **Subnet ID** — The ID for the Spark history server instance. You can use any of the subnets in your VPC. You must be able to reach the network from your client to the subnet. If you want to access via the internet, you must use a public subnet that has the internet gateway in the route table.

   c. Choose **Next**.

4. On the **Configure stack options** page, to use the current user credentials for determining how CloudFormation can create, modify, or delete resources in the stack, choose **Next**. You can also specify a role in the **Permissions** section to use instead of the current user permissions, and then choose **Next**.
5. On the **Review** page, review the template.

   Select **I acknowledge that AWS CloudFormation might create IAM resources**, and then choose **Create stack**.
6. Wait for the stack to be created.
7. Open the **Outputs** tab.

   a. Copy the URL of **SparkUiPublicUrl** if you are using a public subnet.

   b. Copy the URL of **SparkUiPrivateUrl** if you are using a private subnet.
8. Open a web browser, and paste in the URL. This lets you access the server using HTTPS on the specified port. Your browser may not recognize the server's certificate, in which case you have to override its protection and proceed anyway.

Launching the Spark history server and viewing the Spark UI using Docker

If you prefer local access (not to have an EC2 instance for the Apache Spark history server), you can also use Docker to start the Apache Spark history server and view the Spark UI locally. This Dockerfile is a sample that you should modify to meet your requirements.

Prerequisites

For information about how to install Docker on your laptop see the Docker Engine community.

To start the Spark history server and view the Spark UI locally using Docker

1. Download files from GitHub.

Download the Dockerfile and `pom.xml` from [AWS Glue code samples](#).

2. Determine if you want to use your user credentials or federated user credentials to access AWS.

   - To use the current user credentials for accessing AWS, get the values to use for `AWS_ACCESS_KEY_ID` and `AWS_SECRET_ACCESS_KEY` in the `docker run` command. For more information, see [Managing access keys for IAM users](#) in the *IAM User Guide*.
   - To use SAML 2.0 federated users for accessing AWS, get the values for `AWS_ACCESS_KEY_ID`, `AWS_SECRET_ACCESS_KEY`, and `AWS_SESSION_TOKEN`. For more information, see [Requesting temporary security credentials](#)

3. Determine the location of your event log directory, to use in the `docker run` command.

4. Build the Docker image using the files in the local directory, using the name `glue/sparkui`, and the tag `latest`.

   ```
   $ docker build -t glue/sparkui:latest .
   ```

5. Create and start the docker container.

   In the following commands, use the values obtained previously in steps 2 and 3.

   a. To create the docker container using your user credentials, use a command similar to the following

   ```
   docker run -itd -e SPARK_HISTORY_OPTS="$SPARK_HISTORY_OPTS -
   Dspark.history.fs.logDirectory=s3a://path_to_eventlog
    -Dspark.hadoop.fs.s3a.access.key=AWS_ACCESS_KEY_ID -
   Dspark.hadoop.fs.s3a.secret.key=AWS_SECRET_ACCESS_KEY"
    -p 18080:18080 glue/sparkui:latest "/opt/spark/bin/spark-class
    org.apache.spark.deploy.history.HistoryServer"
   ```

   b. To create the docker container using temporary credentials, use `org.apache.hadoop.fs.s3a.TemporaryAWSCredentialsProvider` as the provider, and provide the credential values obtained in step 2. For more information, see [Using Session Credentials with TemporaryAWSCredentialsProvider](#) in the *Hadoop: Integration with Amazon Web Services* documentation.

   ```
   docker run -itd -e SPARK_HISTORY_OPTS="$SPARK_HISTORY_OPTS -
   Dspark.history.fs.logDirectory=s3a://path_to_eventlog
   ```

```
 -Dspark.hadoop.fs.s3a.access.key=AWS_ACCESS_KEY_ID -
Dspark.hadoop.fs.s3a.secret.key=AWS_SECRET_ACCESS_KEY
 -Dspark.hadoop.fs.s3a.session.token=AWS_SESSION_TOKEN
 -
Dspark.hadoop.fs.s3a.aws.credentials.provider=org.apache.hadoop.fs.s3a.TemporaryAWSCred
 -p 18080:18080 glue/sparkui:latest "/opt/spark/bin/spark-class
 org.apache.spark.deploy.history.HistoryServer"
```

> **Note**
>
> These configuration parameters come from the Hadoop-AWS Module. You may need to add specific configuration based on your use case. For example: users in isolated regions will need to configure the `spark.hadoop.fs.s3a.endpoint`.

6. Open `http://localhost:18080` in your browser to view the Spark UI locally.

Monitoring with AWS Glue job run insights

AWS Glue job run insights is a feature in AWS Glue that simplifies job debugging and optimization for your AWS Glue jobs. AWS Glue provides Spark UI, and CloudWatch logs and metrics for monitoring your AWS Glue jobs. With this feature, you get this information about your AWS Glue job's execution:

- Line number of your AWS Glue job script that had a failure.
- Spark action that executed last in the Spark query plan just before the failure of your job.
- Spark exception events related to the failure presented in a time-ordered log stream.
- Root cause analysis and recommended action (such as tuning your script) to fix the issue.
- Common Spark events (log messages relating to a Spark action) with a recommended action that addresses the root cause.

All these insights are available to you using two new log streams in the CloudWatch logs for your AWS Glue jobs.

Requirements

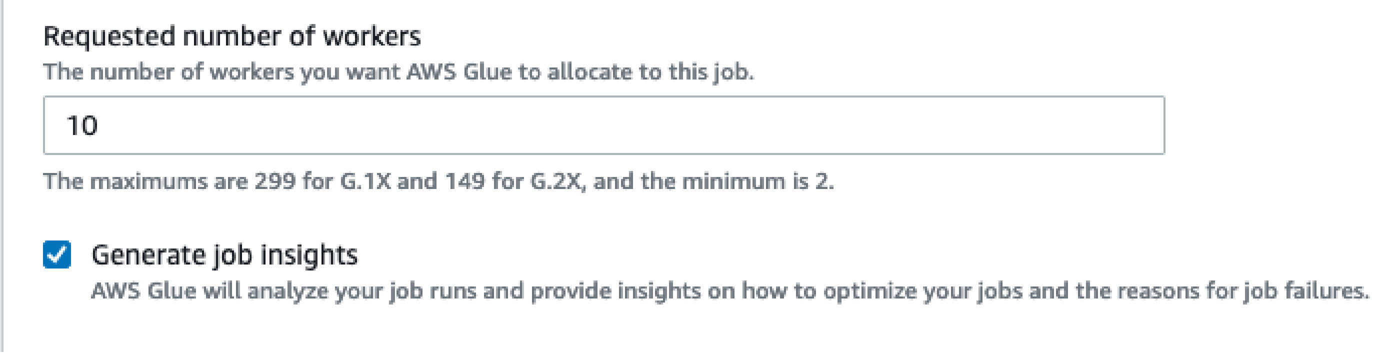
The AWS Glue job run insights feature is available for AWS Glue versions 2.0, 3.0, 4.0, and 5.0. You can follow the migration guide for your existing jobs to upgrade them from older AWS Glue versions.

Enabling job run insights for an AWS Glue ETL job

You can enable job run insights through AWS Glue Studio or the CLI.

AWS Glue Studio

When creating a job via AWS Glue Studio, you can enable or disable job run insights under the **Job Details** tab. Check that the **Generate job insights** box is selected.



Requested number of workers
The number of workers you want AWS Glue to allocate to this job.
10
The maximums are 299 for G.1X and 149 for G.2X, and the minimum is 2.

☑ Generate job insights
AWS Glue will analyze your job runs and provide insights on how to optimize your jobs and the reasons for job failures.

Command line

If creating a job via the CLI, you can start a job run with a single new job parameter: `--enable-job-insights = true`.

By default, the job run insights log streams are created under the same default log group used by AWS Glue continuous logging, that is, `/aws-glue/jobs/logs-v2/`. You may set up custom log group name, log filters and log group configurations using the same set of arguments for continuous logging. For more information, see Enabling Continuous Logging for AWS Glue Jobs.

Accessing the job run insights log streams in CloudWatch

With the job run insights feature enabled, there may be two log streams created when a job run fails. When a job finishes successfully, neither of the streams are generated.

1. *Exception analysis log stream*: `<job-run-id>-job-insights-rca-driver`. This stream provides the following:

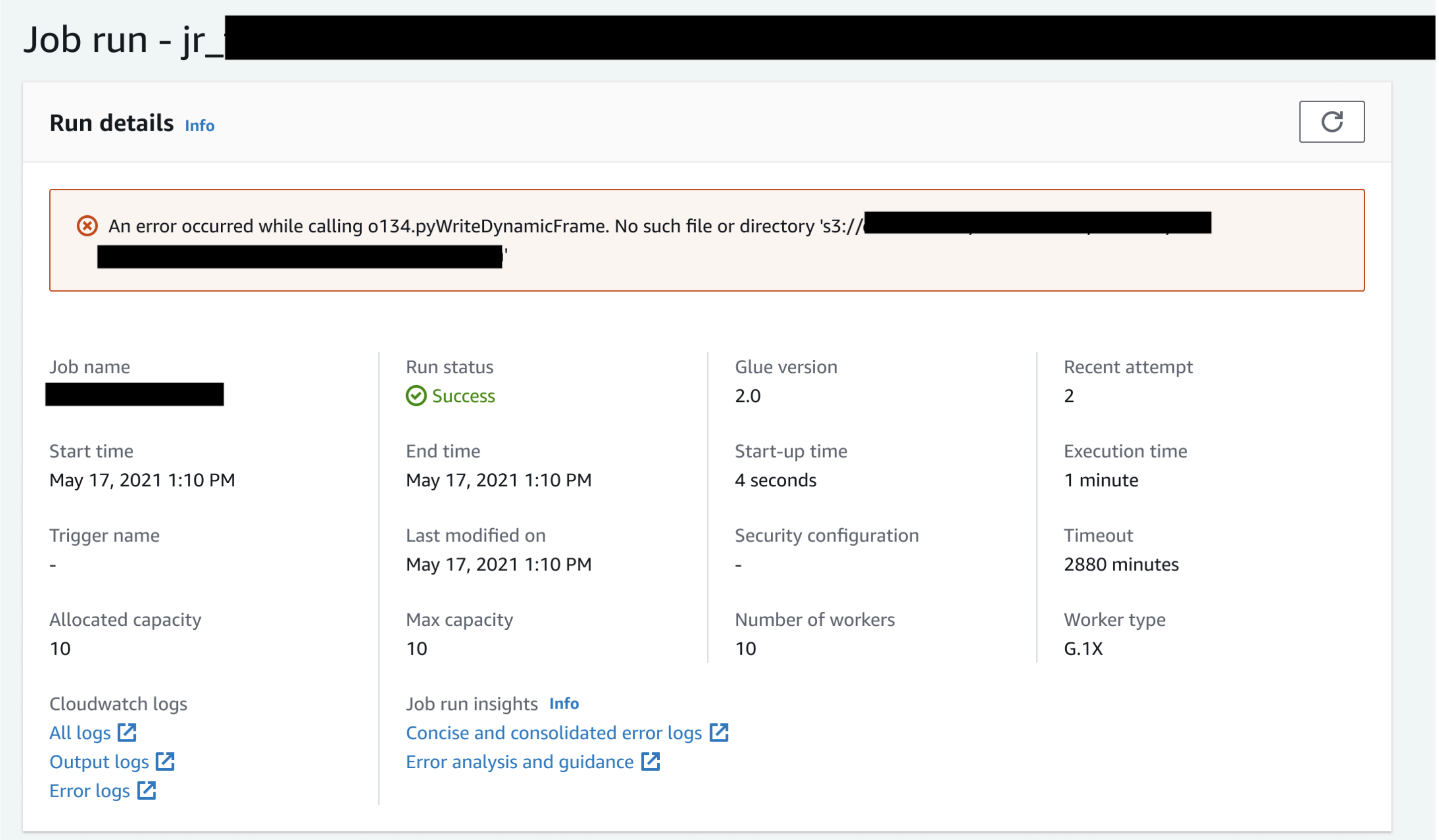
- Line number of your AWS Glue job script that caused the failure.
- Spark action that executed last in the Spark query plan (DAG).
- Concise time-ordered events from the Spark driver and executors that are related to the exception. You can find details such as complete error messages, the failed Spark task and its executors ID that help you to focus on the specific executor's log stream for a deeper investigation if needed.

2. *Rule-based insights stream*:

- Root cause analysis and recommendations on how to fix the errors (such as using a specific job parameter to optimize the performance).
- Relevant Spark events serving as the basis for root cause analysis and a recommended action.

> **Note**
>
> The first stream will exist only if any exception Spark events are available for a failed job run, and the second stream will exist only if any insights are available for the failed job run. For example, if your job finishes successfully, neither of the streams will be generated; if your job fails but there isn't a service-defined rule that can match with your failure scenario, then only the first stream will be generated.

If the job is created from AWS Glue Studio, the links to the above streams are also available under the job run details tab (Job run insights) as "Concise and consolidated error logs" and "Error analysis and guidance".



Job run - jr_[redacted]

Run details Info

An error occurred while calling o134.pyWriteDynamicFrame. No such file or directory 's3://[redacted] [redacted]'

Job name
[redacted]

Run status
Success

Glue version
2.0

Recent attempt
2

Start time
May 17, 2021 1:10 PM

End time
May 17, 2021 1:10 PM

Start-up time
4 seconds

Execution time
1 minute

Trigger name
-

Last modified on
May 17, 2021 1:10 PM

Security configuration
-

Timeout
2880 minutes

Allocated capacity
10

Max capacity
10

Number of workers
10

Worker type
G.1X

Cloudwatch logs
All logs
Output logs
Error logs

Job run insights Info
Concise and consolidated error logs
Error analysis and guidance

Example for AWS Glue job run insights

In this section we present an example of how the job run insights feature can help you resolve an issue in your failed job. In this example, a user forgot to import the required module (tensorflow) in an AWS Glue job to analyze and build a machine learning model on their data.

```
import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job
from pyspark.sql.types import *
from pyspark.sql.functions import udf,col

args = getResolvedOptions(sys.argv, ['JOB_NAME'])

sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
```

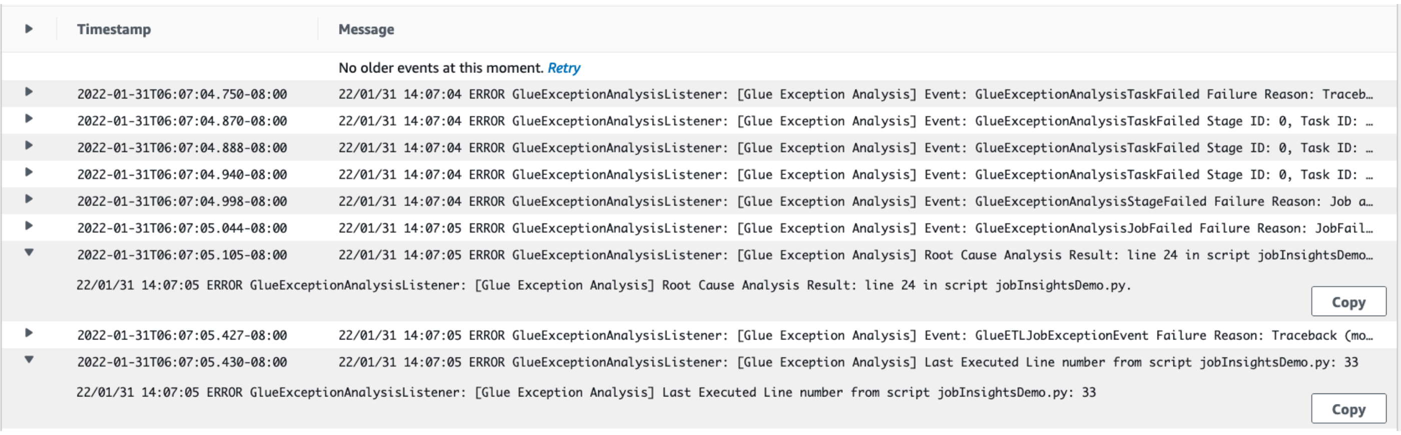
```
job.init(args['JOB_NAME'], args)

data_set_1 = [1, 2, 3, 4]
data_set_2 = [5, 6, 7, 8]

scoresDf = spark.createDataFrame(data_set_1, IntegerType())

def data_multiplier_func(factor, data_vector):
    import tensorflow as tf
    with tf.compat.v1.Session() as sess:
        x1 = tf.constant(factor)
        x2 = tf.constant(data_vector)
        result = tf.multiply(x1, x2)
        return sess.run(result).tolist()

data_multiplier_udf = udf(lambda x:data_multiplier_func(x, data_set_2),
 ArrayType(IntegerType(),False))
factoredDf = scoresDf.withColumn("final_value", data_multiplier_udf(col("value")))
print(factoredDf.collect())
```

Without the job run insights feature, as the job fails, you only see this message thrown by Spark:

```
An error occurred while calling o111.collectToPython. Traceback (most
recent call last):
```

The message is ambiguous and limits your debugging experience. In this case, this feature provides with you additional insights in two CloudWatch log streams:

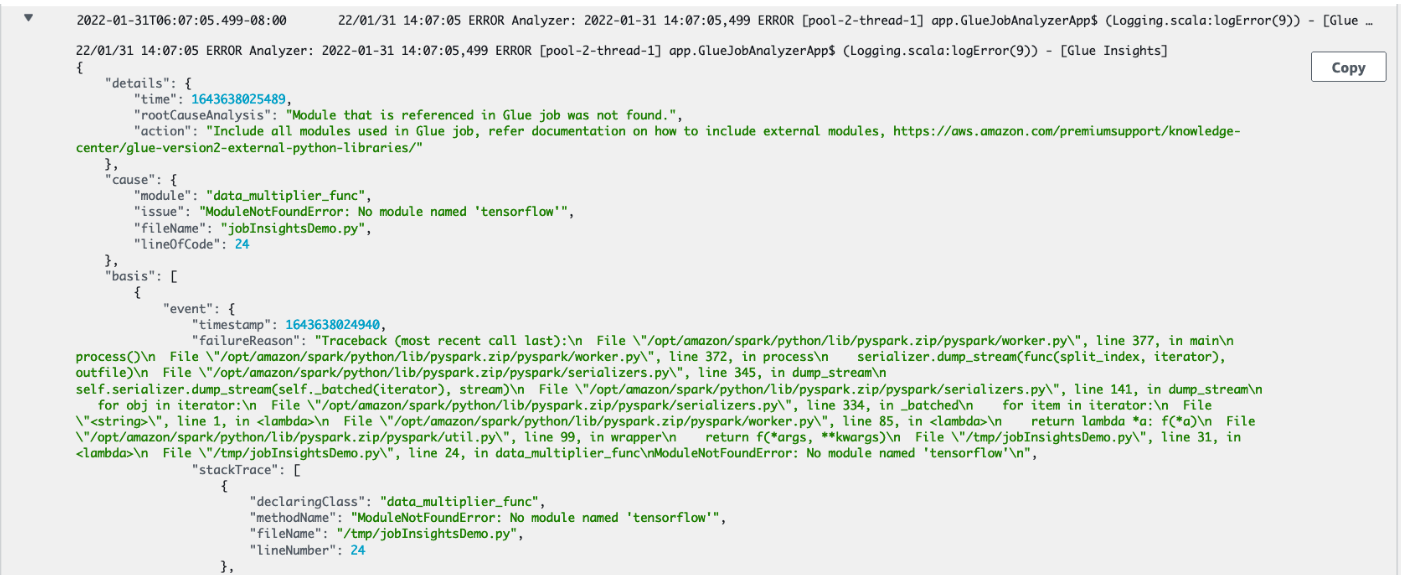
1. The `job-insights-rca-driver` log stream:
    - *Exception events*: This log stream provides you the Spark exception events related to the failure collected from the Spark driver and different distributed workers. These events help you understand the time-ordered propagation of the exception as faulty code executes across Spark tasks, executors, and stages distributed across the AWS Glue workers.
    - *Line numbers*: This log stream identifies line 21, which made the call to import the missing Python module that caused the failure; it also identifies line 24, the call to Spark Action `collect()`, as the last executed line in your script.



```
No older events at this moment. Retry
2022-01-31T06:07:04.750-08:00   22/01/31 14:07:04 ERROR GlueExceptionAnalysisListener: [Glue Exception Analysis] Event: GlueExceptionAnalysisTaskFailed Failure Reason: Traceb…
2022-01-31T06:07:04.870-08:00   22/01/31 14:07:04 ERROR GlueExceptionAnalysisListener: [Glue Exception Analysis] Event: GlueExceptionAnalysisTaskFailed Stage ID: 0, Task ID: …
2022-01-31T06:07:04.888-08:00   22/01/31 14:07:04 ERROR GlueExceptionAnalysisListener: [Glue Exception Analysis] Event: GlueExceptionAnalysisTaskFailed Stage ID: 0, Task ID: …
2022-01-31T06:07:04.940-08:00   22/01/31 14:07:04 ERROR GlueExceptionAnalysisListener: [Glue Exception Analysis] Event: GlueExceptionAnalysisTaskFailed Stage ID: 0, Task ID: …
2022-01-31T06:07:04.998-08:00   22/01/31 14:07:04 ERROR GlueExceptionAnalysisListener: [Glue Exception Analysis] Event: GlueExceptionAnalysisStageFailed Failure Reason: Job a…
2022-01-31T06:07:05.044-08:00   22/01/31 14:07:05 ERROR GlueExceptionAnalysisListener: [Glue Exception Analysis] Event: GlueExceptionAnalysisJobFailed Failure Reason: JobFail…
2022-01-31T06:07:05.105-08:00   22/01/31 14:07:05 ERROR GlueExceptionAnalysisListener: [Glue Exception Analysis] Root Cause Analysis Result: line 24 in script jobInsightsDemo…
22/01/31 14:07:05 ERROR GlueExceptionAnalysisListener: [Glue Exception Analysis] Root Cause Analysis Result: line 24 in script jobInsightsDemo.py.
2022-01-31T06:07:05.427-08:00   22/01/31 14:07:05 ERROR GlueExceptionAnalysisListener: [Glue Exception Analysis] Event: GlueETLJobExceptionEvent Failure Reason: Traceback (mo…
2022-01-31T06:07:05.430-08:00   22/01/31 14:07:05 ERROR GlueExceptionAnalysisListener: [Glue Exception Analysis] Last Executed Line number from script jobInsightsDemo.py: 33
22/01/31 14:07:05 ERROR GlueExceptionAnalysisListener: [Glue Exception Analysis] Last Executed Line number from script jobInsightsDemo.py: 33
```

2. The `job-insights-rule-driver` log stream:

- *Root cause and recommendation*: In addition to the line number and last executed line number for the fault in your script, this log stream shows the root cause analysis and recommendation for you to follow the AWS Glue doc and set up the necessary job parameters in order to use an additional Python module in your AWS Glue job.
- *Basis event*: This log stream also shows the Spark exception event that was evaluated with the service-defined rule to infer the root cause and provide a recommendation.



```
2022-01-31T06:07:05.499-08:00   22/01/31 14:07:05 ERROR Analyzer: 2022-01-31 14:07:05,499 ERROR [pool-2-thread-1] app.GlueJobAnalyzerApp$ (Logging.scala:logError(9)) - [Glue …
22/01/31 14:07:05 ERROR Analyzer: 2022-01-31 14:07:05,499 ERROR [pool-2-thread-1] app.GlueJobAnalyzerApp$ (Logging.scala:logError(9)) - [Glue Insights]
{
    "details": {
        "time": 1643638025489,
        "rootCauseAnalysis": "Module that is referenced in Glue job was not found.",
        "action": "Include all modules used in Glue job, refer documentation on how to include external modules, https://aws.amazon.com/premiumsupport/knowledge-
center/glue-version2-external-python-libraries/"
    },
    "cause": {
        "module": "data_multiplier_func",
        "issue": "ModuleNotFoundError: No module named 'tensorflow'",
        "fileName": "jobInsightsDemo.py",
        "lineOfCode": 24
    },
    "basis": [
        {
            "event": {
                "timestamp": 1643638024940,
                "failureReason": "Traceback (most recent call last):\n  File \"/opt/amazon/spark/python/lib/pyspark.zip/pyspark/worker.py\", line 377, in main\n
process()\n  File \"/opt/amazon/spark/python/lib/pyspark.zip/pyspark/worker.py\", line 372, in process\n    serializer.dump_stream(func(split_index, iterator),
outfile)\n  File \"/opt/amazon/spark/python/lib/pyspark.zip/pyspark/serializers.py\", line 345, in dump_stream\n
self.serializer.dump_stream(self._batched(iterator), stream)\n  File \"/opt/amazon/spark/python/lib/pyspark.zip/pyspark/serializers.py\", line 141, in dump_stream\n
   for obj in iterator:\n  File \"/opt/amazon/spark/python/lib/pyspark.zip/pyspark/serializers.py\", line 334, in _batched\n    for item in iterator:\n  File
\"<string>\", line 1, in <lambda>\n  File \"/opt/amazon/spark/python/lib/pyspark.zip/pyspark/worker.py\", line 85, in <lambda>\n    return lambda *a: f(*a)\n  File
\"/opt/amazon/spark/python/lib/pyspark.zip/pyspark/util.py\", line 99, in wrapper\n    return f(*args, **kwargs)\n  File \"/tmp/jobInsightsDemo.py\", line 31, in
<lambda>\n  File \"/tmp/jobInsightsDemo.py\", line 24, in data_multiplier_func\nModuleNotFoundError: No module named 'tensorflow'\n",
                "stackTrace": [
                    {
                        "declaringClass": "data_multiplier_func",
                        "methodName": "ModuleNotFoundError: No module named 'tensorflow'",
                        "fileName": "/tmp/jobInsightsDemo.py",
                        "lineNumber": 24
                    },
```

Monitoring with Amazon CloudWatch

You can monitor AWS Glue using Amazon CloudWatch, which collects and processes raw data from AWS Glue into readable, near-real-time metrics. These statistics are recorded for a period of two weeks so that you can access historical information for a better perspective on how your web application or service is performing. By default, AWS Glue metrics data is sent to CloudWatch automatically. For more information, see What Is Amazon CloudWatch? in the *Amazon CloudWatch User Guide*, and AWS Glue metrics.

Continous logging

AWS Glue also supports real-time continuous logging for AWS Glue jobs. When continuous logging is enabled for a job, you can view the real-time logs on the AWS Glue console or the CloudWatch console dashboard. For more information, see Logging for AWS Glue jobs.

Observability metrics

When **Job observability metrics** is enabled, additional Amazon CloudWatch metrics are generated when the job is run. Use AWS Glue Observability metrics to generate insights into what is happening inside your AWS Glue to improve triaging and analysis of issues.

Topics

- Monitoring AWS Glue using Amazon CloudWatch metrics
- Setting up Amazon CloudWatch alarms on AWS Glue job profiles
- Logging for AWS Glue jobs
- Monitoring with AWS Glue Observability metrics

Monitoring AWS Glue using Amazon CloudWatch metrics

You can profile and monitor AWS Glue operations using AWS Glue job profiler. It collects and processes raw data from AWS Glue jobs into readable, near real-time metrics stored in Amazon CloudWatch. These statistics are retained and aggregated in CloudWatch so that you can access historical information for a better perspective on how your application is performing.

> **Note**
>
> You may incur additional charges when you enable job metrics and CloudWatch custom metrics are created. For more information, see Amazon CloudWatch pricing .

AWS Glue metrics overview

When you interact with AWS Glue, it sends metrics to CloudWatch. You can view these metrics using the AWS Glue console (the preferred method), the CloudWatch console dashboard, or the AWS Command Line Interface (AWS CLI).

To view metrics using the AWS Glue console dashboard

You can view summary or detailed graphs of metrics for a job, or detailed graphs for a job run.

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, choose **Job run monitoring**.
3. In **Job runs** choose **Actions** to stop a job that is currently running, view a job, or rewind job bookmark.
4. Select a job, then choose **View run details** to view additional information about the job run.

To view metrics using the CloudWatch console dashboard

Metrics are grouped first by the service namespace, and then by the various dimension combinations within each namespace.

1. Open the CloudWatch console at https://console.aws.amazon.com/cloudwatch/.
2. In the navigation pane, choose **Metrics**.
3. Choose the **Glue** namespace.

To view metrics using the AWS CLI

- At a command prompt, use the following command.

  ```
  aws cloudwatch list-metrics --namespace Glue
  ```

AWS Glue reports metrics to CloudWatch every 30 seconds, and the CloudWatch metrics dashboards are configured to display them every minute. The AWS Glue metrics represent delta values from the previously reported values. Where appropriate, metrics dashboards aggregate (sum) the 30-second values to obtain a value for the entire last minute.

AWS Glue metrics behavior for Spark jobs

AWS Glue metrics are enabled at initialization of a `GlueContext` in a script and are generally updated only at the end of an Apache Spark task. They represent the aggregate values across all completed Spark tasks so far.

However, the Spark metrics that AWS Glue passes on to CloudWatch are generally absolute values representing the current state at the time they are reported. AWS Glue reports them to CloudWatch every 30 seconds, and the metrics dashboards generally show the average across the data points received in the last 1 minute.

AWS Glue metrics names are all preceded by one of the following types of prefix:

- `glue.driver.` – Metrics whose names begin with this prefix either represent AWS Glue metrics that are aggregated from all executors at the Spark driver, or Spark metrics corresponding to the Spark driver.
- `glue.`*`executorId`*`.` – The *executorId* is the number of a specific Spark executor. It corresponds with the executors listed in the logs.
- `glue.ALL.` – Metrics whose names begin with this prefix aggregate values from all Spark executors.

AWS Glue metrics

AWS Glue profiles and sends the following metrics to CloudWatch every 30 seconds, and the AWS Glue Metrics Dashboard report them once a minute:

Metric	Description
`glue.driver.aggregate.bytesRead`	The number of bytes read from all data sources by all completed Spark tasks running in all executors. Valid dimensions: `JobName` (the name of the AWS Glue Job), `JobRunId` (the JobRun ID. or `ALL`), and `Type` (count). Valid Statistics: SUM. This metric is a delta value from the last reported value, so on the AWS Glue Metrics Dashboard, a SUM statistic is used for aggregation. Unit: Bytes Can be used to monitor: • Bytes read.

Metric	Description
	- Job progress. - JDBC data sources. - Job Bookmark Issues. - Variance across Job Runs. This metric can be used the same way as the `glue.ALL.s3.filesystem.read_bytes` metric, with the difference that this metric is updated at the end of a Spark task and captures non-S3 data sources as well.
`glue.driver.aggregate.elapsedTime`	The ETL elapsed time in milliseconds (does not include the job bootstrap times). Valid dimensions: `JobName` (the name of the AWS Glue Job), `JobRunId` (the JobRun ID. or `ALL`), and `Type` (count). Valid Statistics: SUM. This metric is a delta value from the last reported value, so on the AWS Glue Metrics Dashboard, a SUM statistic is used for aggregation. Unit: Milliseconds Can be used to determine how long it takes a job run to run on average. Some ways to use the data: - Set alarms for stragglers. - Measure variance across job runs.

Metric	Description
`glue.driver.aggregate.numCompletedStages`	The number of completed stages in the job. Valid dimensions: `JobName` (the name of the AWS Glue Job), `JobRunId` (the JobRun ID. or `ALL`), and `Type` (count). Valid Statistics: SUM. This metric is a delta value from the last reported value, so on the AWS Glue Metrics Dashboard, a SUM statistic is used for aggregation. Unit: Count Can be used to monitor: • Job progress. • Per-stage timeline of job execution,when correlated with other metrics. Some ways to use the data: • Identify demanding stages in the execution of a job. • Set alarms for correlated spikes (demanding stages) across job runs.

Metric	Description
`glue.driver.aggregate.numCompletedTasks`	The number of completed tasks in the job. Valid dimensions: `JobName` (the name of the AWS Glue Job), `JobRunId` (the JobRun ID. or `ALL`), and `Type` (count). Valid Statistics: SUM. This metric is a delta value from the last reported value, so on the AWS Glue Metrics Dashboard, a SUM statistic is used for aggregation. Unit: Count Can be used to monitor: • Job progress. • Parallelism within a stage.

Metric	Description
`glue.driver.aggregate.numFailedTasks`	The number of failed tasks. Valid dimensions: `JobName` (the name of the AWS Glue Job), `JobRunId` (the JobRun ID. or `ALL`), and `Type` (count). Valid Statistics: SUM. This metric is a delta value from the last reported value, so on the AWS Glue Metrics Dashboard, a SUM statistic is used for aggregation. Unit: Count Can be used to monitor: • Data abnormalities that cause job tasks to fail. • Cluster abnormalities that cause job tasks to fail. • Script abnormalities that cause job tasks to fail. The data can be used to set alarms for increased failures that might suggest abnormalities in data, cluster or scripts.

Metric	Description
`glue.driver.aggregate.numKilledTasks`	The number of tasks killed. Valid dimensions: `JobName` (the name of the AWS Glue Job), `JobRunId` (the JobRun ID. or `ALL`), and `Type` (count). Valid Statistics: SUM. This metric is a delta value from the last reported value, so on the AWS Glue Metrics Dashboard, a SUM statistic is used for aggregation. Unit: Count Can be used to monitor: • Abnormalities in Data Skew that result in exceptions (OOMs) that kill tasks. • Script abnormalities that result in exceptions (OOMs) that kill tasks. Some ways to use the data: • Set alarms for increased failures indicating data abnormalities. • Set alarms for increased failures indicating cluster abnormalities. • Set alarms for increased failures indicating script abnormalities.

Metric	Description
`glue.driver.aggregate.recordsRead`	The number of records read from all data sources by all completed Spark tasks running in all executors. Valid dimensions: `JobName` (the name of the AWS Glue Job), `JobRunId` (the JobRun ID. or `ALL`), and `Type` (count). Valid Statistics: SUM. This metric is a delta value from the last reported value, so on the AWS Glue Metrics Dashboard, a SUM statistic is used for aggregation. Unit: Count Can be used to monitor: • Records read. • Job progress. • JDBC data sources. • Job Bookmark Issues. • Skew in Job Runs over days. This metric can be used in a similar way to the `glue.ALL.s3.filesystem.read_bytes` metric, with the difference that this metric is updated at the end of a Spark task.

Metric	Description
`glue.driver.aggregate.shuffleBytesWritten`	The number of bytes written by all executors to shuffle data between them since the previous report (aggregated by the AWS Glue Metrics Dashboard as the number of bytes written for this purpose during the previous minute). Valid dimensions: `JobName` (the name of the AWS Glue Job), `JobRunId` (the JobRun ID. or `ALL`), and `Type` (count). Valid Statistics: SUM. This metric is a delta value from the last reported value, so on the AWS Glue Metrics Dashboard, a SUM statistic is used for aggregation. Unit: Bytes Can be used to monitor: Data shuffle in jobs (large joins, groupBy, repartition, coalesce). Some ways to use the data: • Repartition or decompress large input files before further processing. • Repartition data more uniformly to avoid hot keys. • Pre-filter data before joins or groupBy operations.

Metric	Description
`glue.driver.aggregate.shuffleLocalBytesRead`	The number of bytes read by all executors to shuffle data between them since the previous report (aggregated by the AWS Glue Metrics Dashboard as the number of bytes read for this purpose during the previous minute). Valid dimensions: `JobName` (the name of the AWS Glue Job), `JobRunId` (the JobRun ID. or `ALL`), and `Type` (count). Valid Statistics: SUM. This metric is a delta value from the last reported value, so on the AWS Glue Metrics Dashboard, a SUM statistic is used for aggregation. Unit: Bytes Can be used to monitor: Data shuffle in jobs (large joins, groupBy, repartition, coalesce). Some ways to use the data: • Repartition or decompress large input files before further processing. • Repartition data more uniformly using hot keys. • Pre-filter data before joins or groupBy operations.

Metric	Description
`glue.driver.BlockManager.disk.diskSpaceUsed_MB`	The number of megabytes of disk space used across all executors. Valid dimensions: `JobName` (the name of the AWS Glue Job), `JobRunId` (the JobRun ID. or `ALL`), and `Type` (gauge). Valid Statistics: Average. This is a Spark metric, reported as an absolute value. Unit: Megabytes Can be used to monitor: • Disk space used for blocks that represent cached RDD partitions. • Disk space used for blocks that represent intermediate shuffle outputs. • Disk space used for blocks that represent broadcasts. Some ways to use the data: • Identify job failures due to increased disk usage. • Identify large partitions resulting in spilling or shuffling. • Increase provisioned DPU capacity to correct these issues.

Metric	Description
`glue.driver.ExecutorAllocationManager.executors.numberAllExecutors`	The number of actively running job executors. Valid dimensions: `JobName` (the name of the AWS Glue Job), `JobRunId` (the JobRun ID. or `ALL`), and `Type` (gauge). Valid Statistics: Average. This is a Spark metric, reported as an absolute value. Unit: Count Can be used to monitor: • Job activity. • Straggling executors (with a few executors running only) • Current executor-level parallelism. Some ways to use the data: • Repartition or decompress large input files beforehand if cluster is under-utilized. • Identify stage or job execution delays due to straggler scenarios. • • Compare with numberMaxNeededExecutors to understand backlog for provisioning more DPUs.

Metric	Description
`glue.driver.ExecutorAllocationManager.executors.numberMaxNeededExecutors`	The number of maximum (actively running and pending) job executors needed to satisfy the current load. Valid dimensions: `JobName` (the name of the AWS Glue Job), `JobRunId` (the JobRun ID. or `ALL`), and `Type` (gauge). Valid Statistics: Maximum. This is a Spark metric, reported as an absolute value. Unit: Count Can be used to monitor: • Job activity. • Current executor-level parallelism and backlog of pending tasks not yet scheduled because of unavailable executors due to DPU capacity or killed/failed executors. Some ways to use the data: • Identify pending/backlog of scheduling queue. • Identify stage or job execution delays due to straggler scenarios. • Compare with numberAllExecutors to understand backlog for provisioning more DPUs. • Increase provisioned DPU capacity to correct the pending executor backlog.

Metric	Description

`glue.driver.jvm.heap.usage`
`glue.`*executorId*`.jvm.heap.usage`
`glue.ALL.jvm.heap.usage`
	The fraction of memory used by the JVM heap for this driver (scale: 0-1) for driver, executor identified by executorId, or ALL executors.
	Valid dimensions: `JobName` (the name of the AWS Glue Job), `JobRunId` (the JobRun ID. or `ALL`), and `Type` (gauge).
	Valid Statistics: Average. This is a Spark metric, reported as an absolute value.
	Unit: Percentage
	Can be used to monitor:
	• Driver out-of-memory conditions (OOM) using `glue.driver.jvm.heap.usage` .
	• Executor out-of-memory conditions (OOM) using `glue.ALL.jvm.heap.usage` .
	Some ways to use the data:
	• Identify memory-consuming executor ids and stages.
	• Identify straggling executor ids and stages.
	• Identify a driver out-of-memory condition (OOM).
	• Identify an executor out-of-memory condition (OOM) and obtain the corresponding executor ID so as to be able to get a stack trace from the executor log.
	• Identify files or partitions that may have data skew resulting in stragglers or out-of-memory conditions (OOMs).

Metric	Description

`glue.driver.jvm.heap.used`
`glue.`*executorId*`.jvm.heap.used`
`glue.ALL.jvm.heap.used`
	The number of memory bytes used by the JVM heap for the driver, the executor identified by *executorId*, or ALL executors.
	Valid dimensions: `JobName` (the name of the AWS Glue Job), `JobRunId` (the JobRun ID. or `ALL`), and `Type` (gauge).
	Valid Statistics: Average. This is a Spark metric, reported as an absolute value.
	Unit: Bytes
	Can be used to monitor:
	• Driver out-of-memory conditions (OOM).
	• Executor out-of-memory conditions (OOM).
	Some ways to use the data:
	• Identify memory-consuming executor ids and stages.
	• Identify straggling executor ids and stages.
	• Identify a driver out-of-memory condition (OOM).
	• Identify an executor out-of-memory condition (OOM) and obtain the corresponding executor ID so as to be able to get a stack trace from the executor log.
	• Identify files or partitions that may have data skew resulting in stragglers or out-of-memory conditions (OOMs).

Metric	Description

`glue.driver.s3.filesystem.read_bytes`
`glue.`*executorId*`.s3.filesystem.read_bytes`
`glue.ALL.s3.filesystem.read_bytes`
	The number of bytes read from Amazon S3 by the driver, an executor identified by *executorId*, or ALL executors since the previous report (aggregated by the AWS Glue Metrics Dashboard as the number of bytes read during the previous minute).
	Valid dimensions: `JobName`, `JobRunId`, and `Type` (gauge).
	Valid Statistics: SUM. This metric is a delta value from the last reported value, so on the AWS Glue Metrics Dashboard a SUM statistic is used for aggregation. The area under the curve on the AWS Glue Metrics Dashboard can be used to visually compare bytes read by two different job runs.
	Unit: Bytes.
	Can be used to monitor:
	• ETL data movement.
	• Job progress.
	• Job bookmark issues (data processed, reprocessed, and skipped).
	• Comparison of reads to ingestion rate from external data sources.
	• Variance across job runs.
	Resulting data can be used for:
	• DPU capacity planning.
	• Setting alarms for large spikes or dips in data read for job runs and job stages.

Metric	Description

`glue.driver.s3.filesystem.write_bytes`
`glue.`*executorId*`.s3.filesystem.write_bytes`
`glue.ALL.s3.filesystem.write_bytes`
	The number of bytes written to Amazon S3 by the driver, an executor identified by *executorId*, or ALL executors since the previous report (aggregated by the AWS Glue Metrics Dashboard as the number of bytes written during the previous minute).
	Valid dimensions: `JobName`, `JobRunId`, and `Type` (gauge).
	Valid Statistics: SUM. This metric is a delta value from the last reported value, so on the AWS Glue Metrics Dashboard a SUM statistic is used for aggregation. The area under the curve on the AWS Glue Metrics Dashboard can be used to visually compare bytes written by two different job runs.
	Unit: Bytes
	Can be used to monitor:
	• ETL data movement.
	• Job progress.
	• Job bookmark issues (data processed, reprocessed, and skipped).
	• Comparison of reads to ingestion rate from external data sources.
	• Variance across job runs.
	Some ways to use the data:
	• DPU capacity planning.
	• Setting alarms for large spikes or dips in data read for job runs and job stages.

Metric	Description
`glue.driver.streaming.numRecords`	The number of records that are received in a micro-batch. This metric is only available for AWS Glue streaming jobs with AWS Glue version 2.0 and above. Valid dimensions: `JobName` (the name of the AWS Glue job), `JobRunId` (the JobRun ID. or `ALL`), and `Type` (count). Valid Statistics: Sum, Maximum, Minimum, Average, Percentile Unit: Count Can be used to monitor: • Records read. • Job progress.
`glue.driver.streaming.batchProcessingTimeInMs`	The time it takes to process the batches in milliseconds. This metric is only available for AWS Glue streaming jobs with AWS Glue version 2.0 and above. Valid dimensions: `JobName` (the name of the AWS Glue job), `JobRunId` (the JobRun ID. or `ALL`), and `Type` (count). Valid Statistics: Sum, Maximum, Minimum, Average, Percentile Unit: Count Can be used to monitor: • Job progress. • Script performance.

Metric	Description

`glue.driver.system.cpuSystemLoad`
`glue.`*executorId*`.system.cpuSystemLoad`
`glue.ALL.system.cpuSystemLoad`
	The fraction of CPU system load used (scale: 0-1) by the driver, an executor identified by *executorId*, or ALL executors.
	Valid dimensions: `JobName` (the name of the AWS Glue job), `JobRunId` (the JobRun ID. or `ALL`), and `Type` (gauge).
	Valid Statistics: Average. This metric is reported as an absolute value.
	Unit: Percentage
	Can be used to monitor:
	• Driver CPU load.
	• Executor CPU load.
	• Detecting CPU-bound or IO-bound executors or stages in a Job.
	Some ways to use the data:
	• DPU capacity Planning along with IO Metrics (Bytes Read/Shuffle Bytes, Task Parallelism) and the number of maximum needed executors metric.
	• Identify the CPU/IO-bound ratio. This allows for repartitionioning and increasing provisioned capacity for long-running jobs with splittable datasets having lower CPU utilization.

Dimensions for AWS Glue Metrics

AWS Glue metrics use the AWS Glue namespace and provide metrics for the following dimensions:

Dimension	Description
`JobName`	This dimension filters for metrics of all job runs of a specific AWS Glue job.
`JobRunId`	This dimension filters for metrics of a specific AWS Glue job run by a JobRun ID, or `ALL`.
`Type`	This dimension filters for metrics by either `count` (an aggregate number) or `gauge` (a value at a point in time).

For more information, see the Amazon CloudWatch User Guide.

Setting up Amazon CloudWatch alarms on AWS Glue job profiles

AWS Glue metrics are also available in Amazon CloudWatch. You can set up alarms on any AWS Glue metric for scheduled jobs.

A few common scenarios for setting up alarms are as follows:

- Jobs running out of memory (OOM): Set an alarm when the memory usage exceeds the normal average for either the driver or an executor for an AWS Glue job.
- Straggling executors: Set an alarm when the number of executors falls below a certain threshold for a large duration of time in an AWS Glue job.
- Data backlog or reprocessing: Compare the metrics from individual jobs in a workflow using a CloudWatch math expression. You can then trigger an alarm on the resulting expression value (such as the ratio of bytes written by a job and bytes read by a following job).

For detailed instructions on setting alarms, see Create or Edit a CloudWatch Alarm in the *Amazon CloudWatch Events User Guide*.

For monitoring and debugging scenarios using CloudWatch, see Job monitoring and debugging.

Logging for AWS Glue jobs

In AWS Glue 5.0, all jobs have real-time logging capabilities. Additionally, you can specify custom configuration options to tailor the logging behavior. These options include setting the Amazon

CloudWatch log group name, the Amazon CloudWatch log stream prefix (which will precede the AWS Glue job run ID and driver/executor ID), and the log conversion pattern for log messages. These configurations allow you to aggregate logs in custom Amazon CloudWatch log groups with different expiration policies. Furthermore, you can analyze the logs more effectively by using custom log stream prefixes and conversion patterns. This level of customization enables you to optimize log management and analysis according to your specific requirements.

Logging behavior in AWS Glue 5.0

By default, system logs, Spark daemon logs, and user AWS Glue Logger logs are written to the `/aws-glue/jobs/error` log group in Amazon CloudWatch. On the other hand, user stdout (standard output) and stderr (standard error) logs are written to the `/aws-glue/jobs/output` log group by default.

Custom logging

You can customize the default log group and log stream prefixes using the following job arguments:

- `--custom-logGroup-prefix`: Allows you to specify a custom prefix for the `/aws-glue/jobs/error` and `/aws-glue/jobs/output` log groups. If you provide a custom prefix, the log group names will be in the following format:
  - `/aws-glue/jobs/error` will be `<customer prefix>/error`
  - `/aws-glue/jobs/output` will be `<customer prefix>/output`
- `--custom-logStream-prefix`: Allows you to specify a custom prefix for the log stream names within the log groups. If you provide a custom prefix, the log stream names will be in the following format:
  - `jobrunid-driver` will be `<customer log stream>-driver`
  - `jobrunid-executorNum` will be `<customer log stream>-executorNum`

Validation rules and limitations for custom prefixes:

- The entire log stream name must be between 1 and 512 characters long.
- The custom prefix itself is restricted to 400 characters.
- The custom prefix must match the regular expression pattern `[^:*]*` (special characters allowed are '_', '-', and '/').

Logging application-specific messages using the custom script logger

You can use the AWS Glue logger to log any application-specific messages in the script that are sent in real time to the driver log stream.

The following example shows a Python script.

```
from awsglue.context import GlueContext
from pyspark.context import SparkContext

sc = SparkContext()
glueContext = GlueContext(sc)
logger = glueContext.get_logger()
logger.info("info message")
logger.warn("warn message")
logger.error("error message")
```

The following example shows a Scala script.

```
import com.amazonaws.services.glue.log.GlueLogger

object GlueApp {
  def main(sysArgs: Array[String]) {
    val logger = new GlueLogger
    logger.info("info message")
    logger.warn("warn message")
    logger.error("error message")
  }
}
```

Enabling the progress bar to show job progress

AWS Glue provides a real-time progress bar under the `JOB_RUN_ID-progress-bar` log stream to check AWS Glue job run status. Currently it supports only jobs that initialize `glueContext`. If you run a pure Spark job without initializing `glueContext`, the AWS Glue progress bar does not appear.

The progress bar shows the following progress update every 5 seconds.

```
Stage Number (Stage Name): > (numCompletedTasks + numActiveTasks) /
 totalNumOfTasksInThisStage]
```

Security configuration with Amazon CloudWatch logging

When a security configuration is enabled for Amazon CloudWatch logs, AWS Glue creates log groups with specific naming patterns that incorporate the security configuration name.

Log group naming with security configuration

The default and custom log groups will be as follows:

- **Default error log group:** `/aws-glue/jobs/`*`Security-Configuration-Name`*`-role/`*`glue-job-role`*`/error`
- **Default output log group:** `/aws-glue/jobs/`*`Security-Configuration-Name`*`-role/`*`glue-job-role`*`/output`
- **Custom error log group (AWS Glue 5.0):** *`custom-log-group-prefix`*`/`*`Security-Configuration-Name`*`-role/`*`glue-job-role`*`/error`
- **Custom output log group (AWS Glue 5.0):** *`custom-log-group-prefix`*`/`*`Security-Configuration-Name`*`-role/`*`glue-job-role`*`/output`

Required IAM Permissions

You need to add the `logs:AssociateKmsKey` permission to your IAM role permissions, if you enable a security configuration with Amazon CloudWatch Logs. If that permission is not included, continuous logging will be disabled.

Also, to configure the encryption for the Amazon CloudWatch Logs, follow the instructions at Encrypt Log Data in Amazon CloudWatch Logs Using AWS Key Management Service in the Amazon Amazon CloudWatch Logs User Guide.

Additional Information

For more information on creating security configurations, see Managing security configurations on the AWS Glue console.

Topics

- Enabling continuous logging for AWS Glue 4.0 and earlier jobs
- Viewing logs for AWS Glue jobs

Enabling continuous logging for AWS Glue 4.0 and earlier jobs

> **Note**
>
> In AWS Glue 4.0 and earlier versions, continuous logging was an available feature. However, with the introduction of AWS Glue 5.0, all jobs have real-time logging capability. For more details on the logging capabilities and configuration options in AWS Glue 5.0, see Logging for AWS Glue jobs.

You can enable continuous logging using the AWS Glue console or through the AWS Command Line Interface (AWS CLI).

You can enable continuous logging when you create a new job, edit an existing job, or enable it through the AWS CLI.

You can also specify custom configuration options such as the Amazon CloudWatch log group name, CloudWatch log stream prefix before the AWS Glue job run ID driver/executor ID, and log conversion pattern for log messages. These configurations help you to set aggregate logs in custom CloudWatch log groups with different expiration policies, and analyze them further with custom log stream prefixes and conversions patterns.

Topics

- Using the AWS Management Console
- Logging application-specific messages using the custom script logger
- Enabling the progress bar to show job progress
- Security configuration with continuous logging

Using the AWS Management Console

Follow these steps to use the console to enable continuous logging when creating or editing an AWS Glue job.

To create a new AWS Glue job with continuous logging

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, choose **ETL jobs**.

3. Choose **Visual ETL**.
4. In the **Job details** tab, expand the **Advanced properties** section.
5. Under **Continuous logging** select **Enable logs in CloudWatch**.

To enable continuous logging for an existing AWS Glue job

1. Open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, choose **Jobs**.
3. Choose an existing job from the **Jobs** list.
4. Choose **Action**, **Edit job**.
5. In the **Job details** tab, expand the **Advanced properties** section.
6. Under **Continuous logging** select **Enable logs in CloudWatch**.

Using the AWS CLI

To enable continuous logging, you pass in job parameters to an AWS Glue job. Pass the following special job parameters similar to other AWS Glue job parameters. For more information, see Using job parameters in AWS Glue jobs.

```
'--enable-continuous-cloudwatch-log': 'true'
```

You can specify a custom Amazon CloudWatch log group name. If not specified, the default log group name is `/aws-glue/jobs/logs-v2`.

```
'--continuous-log-logGroup': 'custom_log_group_name'
```

You can specify a custom Amazon CloudWatch log stream prefix. If not specified, the default log stream prefix is the job run ID.

```
'--continuous-log-logStreamPrefix': 'custom_log_stream_prefix'
```

You can specify a custom continuous logging conversion pattern. If not specified, the default conversion pattern is `%d{yy/MM/dd HH:mm:ss} %p %c{1}: %m%n`. Note that the conversion pattern only applies to driver logs and executor logs. It does not affect the AWS Glue progress bar.

```
'--continuous-log-conversionPattern': 'custom_log_conversion_pattern'
```

Logging application-specific messages using the custom script logger

You can use the AWS Glue logger to log any application-specific messages in the script that are sent in real time to the driver log stream.

The following example shows a Python script.

```
from awsglue.context import GlueContext
from pyspark.context import SparkContext

sc = SparkContext()
glueContext = GlueContext(sc)
logger = glueContext.get_logger()
logger.info("info message")
logger.warn("warn message")
logger.error("error message")
```

The following example shows a Scala script.

```
import com.amazonaws.services.glue.log.GlueLogger

object GlueApp {
  def main(sysArgs: Array[String]) {
    val logger = new GlueLogger
    logger.info("info message")
    logger.warn("warn message")
    logger.error("error message")
  }
}
```

Enabling the progress bar to show job progress

AWS Glue provides a real-time progress bar under the `JOB_RUN_ID-progress-bar` log stream to check AWS Glue job run status. Currently it supports only jobs that initialize `glueContext`. If you run a pure Spark job without initializing `glueContext`, the AWS Glue progress bar does not appear.

The progress bar shows the following progress update every 5 seconds.

```
Stage Number (Stage Name): > (numCompletedTasks + numActiveTasks) /
 totalNumOfTasksInThisStage]
```

Security configuration with continuous logging

If a security configuration is enabled for CloudWatch logs, AWS Glue will create a log group named as follows for continuous logs:

```
<Log-Group-Name>-<Security-Configuration-Name>
```

The default and custom log groups will be as follows:

- The default continuous log group will be `/aws-glue/jobs/error-<Security-Configuration-Name>`
- The custom continuous log group will be `<custom-log-group-name>-<Security-Configuration-Name>`

You need to add the `logs:AssociateKmsKey` to your IAM role permissions, if you enable a security configuration with CloudWatch Logs. If that permission is not included, continuous logging will be disabled. Also, to configure the encryption for the CloudWatch Logs, follow the instructions at Encrypt Log Data in CloudWatch Logs Using AWS Key Management Service in the *Amazon CloudWatch Logs User Guide*.

For more information on creating security configurations, see Managing security configurations on the AWS Glue console.

> **Note**
>
> You may incur additional charges when you enable logging and additional CloudWatch log events are created. For more information, see Amazon CloudWatch pricing .

Viewing logs for AWS Glue jobs

You can view real-time logs using the AWS Glue console or the Amazon CloudWatch console.

To view real-time logs using the AWS Glue console dashboard

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, choose **Jobs**.

3. Add or start an existing job. Choose **Action**, **Run job**.

   When you start running a job, you navigate to a page that contains information about the running job:

   - The **Logs** tab shows the older aggregated application logs.
   - The **Logs** tab shows a real-time progress bar when the job is running with `glueContext` initialized.
   - The **Logs** tab also contains the **Driver logs**, which capture real-time Apache Spark driver logs, and application logs from the script logged using the AWS Glue application logger when the job is running.
4. For older jobs, you can also view the real-time logs under the **Job History** view by choosing **Logs**. This action takes you to the CloudWatch console that shows all Spark driver, executor, and progress bar log streams for that job run.

To view real-time logs using the CloudWatch console dashboard

1. Open the CloudWatch console at https://console.aws.amazon.com/cloudwatch/.
2. In the navigation pane, choose **Log**.
3. Choose the **/aws-glue/jobs/error/** log group.
4. In the **Filter** box, paste the job run ID.

   You can view the driver logs, executor logs, and progress bar (if using the **Standard filter**).

Monitoring with AWS Glue Observability metrics

> **Note**
> AWS Glue Observability metrics is available on AWS Glue 4.0 and later versions.

Use AWS Glue Observability metrics to generate insights into what is happening inside your AWS Glue for Apache Spark jobs to improve triaging and analysis of issues. Observability metrics are visualized through Amazon CloudWatch dashboards and can be used to help perform root cause analysis for errors and for diagnosing performance bottlenecks. You can reduce the time spent debugging issues at scale so you can focus on resolving issues faster and more effectively.

AWS Glue Observability provides Amazon CloudWatch metrics categorized in following four groups:

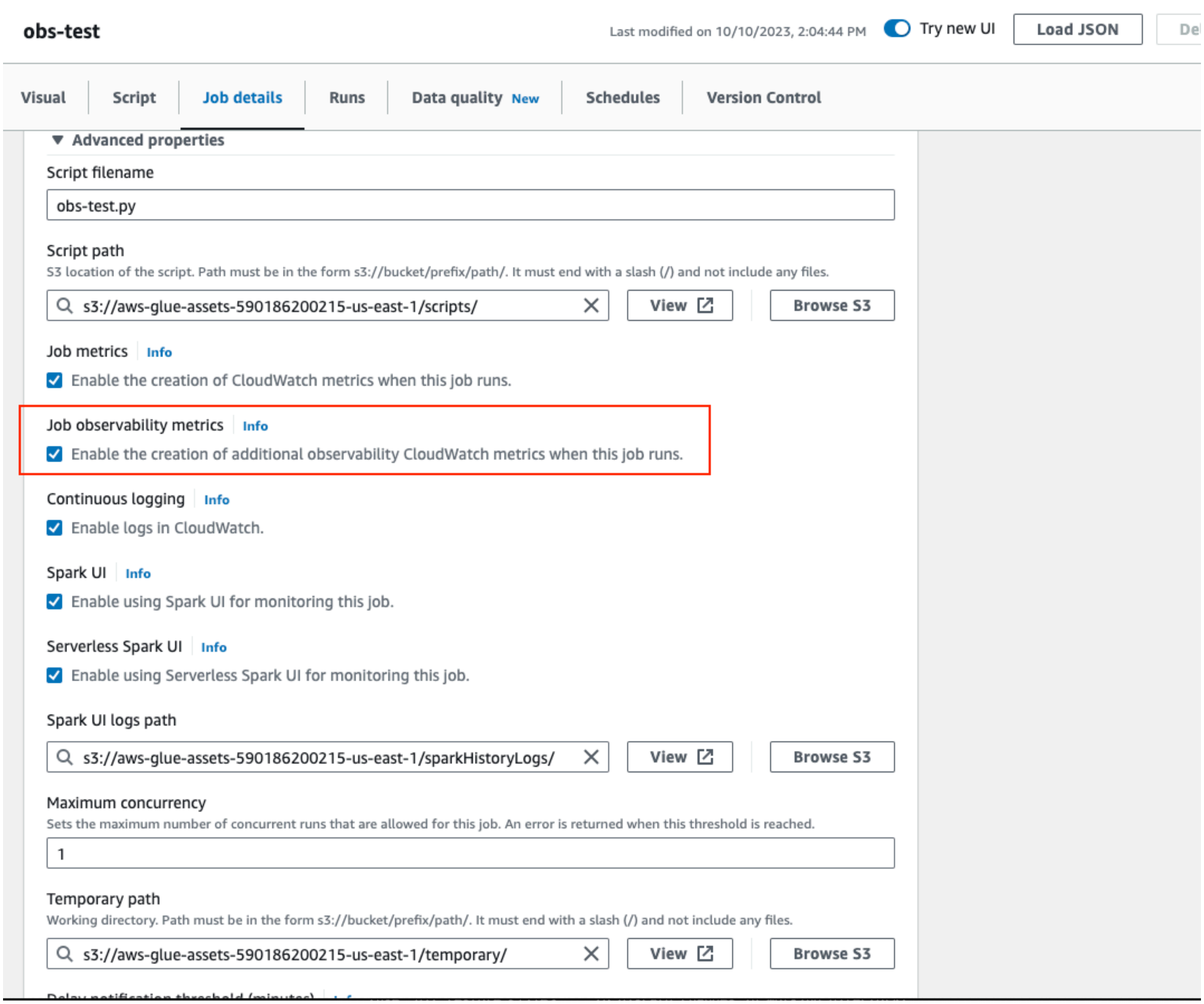
- **Reliability (i.e., Errors Classes) –** easily identify the most common failure reasons at given time range that you may want to address.
- **Performance (i.e., Skewness) –** identify a performance bottleneck and apply tuning techniques. For example, when you experience degraded performance due to job skewness, you may want to enable Spark Adaptive Query Execution and fine-tune the skew join threshold.
- **Throughput (i.e., per source/sink throughput) –** monitor trends of data reads and writes. You can also configure Amazon CloudWatch alarms for anomalies.
- **Resource Utilization (i.e., workers, memory and disk utilization) –** efficiently find the jobs with low capacity utilization. You may want to enable AWS Glue auto-scaling for those jobs.

Getting started with AWS Glue Observability metrics

> **Note**
>
> The new metrics are enabled by default in the AWS Glue Studio console.

To configure observability metrics in AWS Glue Studio:

1. Log in to the AWS Glue console and choose **ETL jobs** from the console menu.
2. Choose a job by clicking on the job name in the **Your jobs** section.
3. Choose the **Job details** tab.
4. Scroll to the bottom and choose **Advanced properties**, then **Job observability metrics**.



obs-test

Last modified on 10/10/2023, 2:04:44 PM

▼ Advanced properties

Script filename

obs-test.py

Script path

S3 location of the script. Path must be in the form s3://bucket/prefix/path/. It must end with a slash (/) and not include any files.

s3://aws-glue-assets-590186200215-us-east-1/scripts/

Job metrics

☑ Enable the creation of CloudWatch metrics when this job runs.

Job observability metrics

☑ Enable the creation of additional observability CloudWatch metrics when this job runs.

Continuous logging

☑ Enable logs in CloudWatch.

Spark UI

☑ Enable using Spark UI for monitoring this job.

Serverless Spark UI

☑ Enable using Serverless Spark UI for monitoring this job.

Spark UI logs path

s3://aws-glue-assets-590186200215-us-east-1/sparkHistoryLogs/

Maximum concurrency

Sets the maximum number of concurrent runs that are allowed for this job. An error is returned when this threshold is reached.

1

Temporary path

Working directory. Path must be in the form s3://bucket/prefix/path/. It must end with a slash (/) and not include any files.

s3://aws-glue-assets-590186200215-us-east-1/temporary/

To enable AWS Glue Observability metrics using AWS CLI:

- Add to the `--default-arguments` map the following key-value in the input JSON file:

  ```
  --enable-observability-metrics, true
  ```

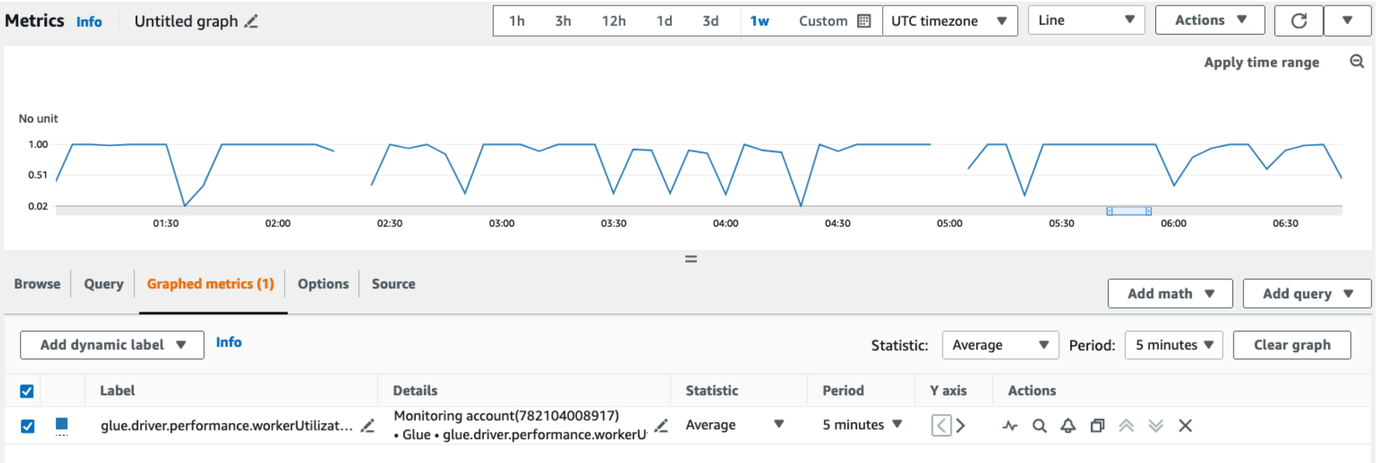
Using AWS Glue observability

Because the AWS Glue observability metrics is provided through Amazon CloudWatch, you can use the Amazon CloudWatch console, AWS CLI, SDK or API to query the observability metrics datapoints. See [Using Glue Observability for monitoring resource utilization to reduce cost](#) for an example use case when to use AWS Glue observability metrics.

Using AWS Glue observability in the Amazon CloudWatch console

To query and visualize metrics in the Amazon CloudWatch console:

1. Open the Amazon CloudWatch console and choose **All metrics**.
2. Under custom namespaces, choose **AWS Glue**.
3. Choose **Job Observability Metrics, Observability Metrics Per Source, or Observability Metrics Per Sink** .
4. Search for the specific metric name, job name, job run ID, and select them.
5. Under the **Graphed metrics** tab, configure your preferred statistic, period, and other options.


Metrics Info
Untitled graph
1h
3h
12h
1d
3d
1w
Custom
UTC timezone
Line
Actions
Apply time range
No unit
1.00
0.51
0.02
01:30
02:00
02:30
03:00
03:30
04:00
04:30
05:00
05:30
06:00
06:30
Browse
Query
Graphed metrics (1)
Options
Source
Add math
Add query
Add dynamic label
Info
Clear graph

Statistic: Average
Period: 5 minutes

☑		Label	Details	Statistic	Period	Y axis	Actions
☑	🟦	glue.driver.performance.workerUtilizat...	Monitoring account(782104008917) • Glue • glue.driver.performance.workerU	Average	5 minutes		

To query an Observability metric using AWS CLI:

1. Create a metric definition JSON file and replace `your-Glue-job-name`and `your-Glue-job-run-id` with yours.

```
$ cat multiplequeries.json
[
    {
        "Id": "avgWorkerUtil_0",
        "MetricStat": {
            "Metric": {
                "Namespace": "Glue",
                "MetricName": "glue.driver.workerUtilization",
                "Dimensions": [
                    {
                        "Name": "JobName",
                        "Value": "<your-Glue-job-name-A>"
                    },
                    {
                        "Name": "JobRunId",
                        "Value": "<your-Glue-job-run-id-A>"
                    },
                    {
                        "Name": "Type",
                        "Value": "gauge"
                    },
                    {
                        "Name": "ObservabilityGroup",
                        "Value": "resource_utilization"
                    }
                ]
            },
            "Period": 1800,
            "Stat": "Minimum",
            "Unit": "None"
        }
    },
    {
        "Id": "avgWorkerUtil_1",
        "MetricStat": {
            "Metric": {
                "Namespace": "Glue",
                "MetricName": "glue.driver.workerUtilization",
                "Dimensions": [
                    {
                        "Name": "JobName",
                        "Value": "<your-Glue-job-name-B>"
                    },
                    {
                        "Name": "JobRunId",
                        "Value": "<your-Glue-job-run-id-B>"
                    },
                    {
                        "Name": "Type",
                        "Value": "gauge"
                    },
                    {
                        "Name": "ObservabilityGroup",
                        "Value": "resource_utilization"
                    }
                ]
            },
            "Period": 1800,
            "Stat": "Minimum",
            "Unit": "None"
        }
    }
]
```

2. Run the `get-metric-data` command:

```
$ aws cloudwatch get-metric-data --metric-data-queries file: //multiplequeries.json
 \
     --start-time '2023-10-28T18: 20' \
     --end-time '2023-10-28T19: 10'  \
     --region us-east-1
{
    "MetricDataResults": [
        {
            "Id": "avgWorkerUtil_0",
            "Label": "<your-label-for-A>",
            "Timestamps": [
                "2023-10-28T18:20:00+00:00"
            ],
            "Values": [
                0.06718750000000001
            ],
            "StatusCode": "Complete"
        },
        {
            "Id": "avgWorkerUtil_1",
            "Label": "<your-label-for-B>",
            "Timestamps": [
                "2023-10-28T18:50:00+00:00"
            ],
            "Values": [
                0.5959183673469387
            ],
            "StatusCode": "Complete"
        }
    ],
    "Messages": []
}
```

Observability metrics

AWS Glue Observability profiles and sends the following metrics to Amazon CloudWatch every 30 seconds, and some of these metrics can be visible in the AWS Glue Studio Job Runs Monitoring Page.

Metric	Description	Category
glue.driver.skewness.stage	Metric Category: job_performance The spark stages execution Skewness: this metric captures execution skewness, which might be caused by input data skewness or by a transformation (e.g., skewed join). The values of this metric falls into the range of [0, infinity[, where 0 means the ratio of the maximum to median tasks' execution time, among all tasks in the stage is less than a certain stage skewness factor. The default stage skewness factor is `5` and it be overwritten via spark conf: spark.metrics.conf.driver.source.glue.jobPerformance.skewnessFactor	job_performance

Metric	Description	Category
	A stage skewness value of 1 means the ratio is twice the stage skewness factor. The value of stage skewness is updated every 30 seconds to reflect the current skewness. The value at the end of the stage reflects the final stage skewness. Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (gauge), and ObservabilityGroup (job_performance) Valid Statistics: Average, Maximum, Minimum, Percentile Unit: Count	

Metric	Description	Category
glue.driver.skewness.job	Metric Category: job_performance Job skewness is the weighted average of the job stages skewness. Weighted average gives more weight to stages that takes longer to execute. This is to avoid the corner case when a very skewwed stage is actually running for very short time relative to other stages (and thus its skewness is not significant for the overall job perforamnce and does not worth the effort to try to address its skewness) . This metric is updated upon completion of each stage, and thus the last value reflects the actual overall job skewness. Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (gauge), and ObservabilityGroup (job_performance) Valid Statistics: Average, Maximum, Minimum, Percentile Unit: Count	job_performance

Metric	Description	Category
glue.succeed.ALL	Metric Category: error Total number of successful job runs, to complete the picture of failures categories Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (count), and ObservabilityGroup (error) Valid Statistics: SUM Unit: Count	error
glue.error.ALL	Metric Category: error Total number of job run errors, to complete the picture of failures categories Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (count), and ObservabilityGroup (error) Valid Statistics: SUM Unit: Count	error

Metric	Description	Category
glue.error.[error category]	Metric Category: error This is actually a set of metrics, that are updated only when a job run fails. The error categorization helps with triaging and debugging. When a job run fails, the error causing the failure is categorized and the corresponding error category metric is set to 1. This helps to perform over time failures analysis, as well as over all jobs error analysis to identify most common failure categories to start addressing them. AWS Glue has 28 error categories, including OUT_OF_MEMORY (driver and executor), PERMISSION, SYNTAX and THROTTLING error categories. Error categories also include COMPILATION, LAUNCH and TIMEOUT error categories. Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (count), and ObservabilityGroup (error) Valid Statistics: SUM	error

Metric	Description	Category
	Unit: Count	
glue.driver.workerUtilization	Metric Category: resource_ utilization The percentage of the allocated workers which are actually used. If not good, auto scaling can help. Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (gauge), and ObservabilityGroup (resource _utilization) Valid Statistics: Average, Maximum, Minimum, Percentile Unit: Percentage	resource_utilization

Metric	Description	Category
glue.driver.memory.heap.[available | used]	Metric Category: resource_utilization The driver's available / used heap memory during the job run. This helps to understand memory usage trends, especially over time, which can help avoid potential failures, in addition to debugging memory related failures. Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (gauge), and ObservabilityGroup (resource_utilization) Valid Statistics: Average Unit: Bytes	resource_utilization

Metric	Description	Category
glue.driver.memory.heap.used.percentage	Metric Category: resource_utilization The driver's used (%) heap memory during the job run. This helps to understand memory usage trends, especially over time, which can help avoid potential failures, in addition to debugging memory related failures. Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (gauge), and ObservabilityGroup (resource_utilization) Valid Statistics: Average Unit: Percentage	resource_utilization

Metric	Description	Category
glue.driver.memory.non-heap.[available | used]	Metric Category: resource_utilization The driver's available / used non-heap memory during the job run. This helps to understand memory usage trensd, especially over time, which can help avoid potential failures, in addition to debugging memory related failures. Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (gauge), and ObservabilityGroup (resource_utilization) Valid Statistics: Average Unit: Bytes	resource_utilization

Metric	Description	Category
glue.driver.memory.non-heap.used.percentage	Metric Category: resource_utilization The driver's used (%) non-heap memory during the job run. This helps to understand memory usage trends, especially over time, which can help avoid potential failures, in addition to debugging memory related failures. Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (gauge), and ObservabilityGroup (resource_utilization) Valid Statistics: Average Unit: Percentage	resource_utilization

Metric	Description	Category
glue.driver.memory.total.[available | used]	Metric Category: resource_utilization The driver's available / used total memory during the job run. This helps to understand memory usage trends, especially over time, which can help avoid potential failures, in addition to debugging memory related failures. Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (gauge), and ObservabilityGroup (resource_utilization) Valid Statistics: Average Unit: Bytes	resource_utilization

Metric	Description	Category
glue.driver.memory.total.used.percentage	Metric Category: resource_utilization The driver's used (%) total memory during the job run. This helps to understand memory usage trends, especially over time, which can help avoid potential failures, in addition to debugging memory related failures. Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (gauge), and ObservabilityGroup (resource_utilization) Valid Statistics: Average Unit: Percentage	resource_utilization

Metric	Description	Category
glue.ALL.memory.heap.[available | used]	Metric Category: resource_utilization The executors' available/used heap memory. ALL means all executors. Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (gauge), and ObservabilityGroup (resource_utilization) Valid Statistics: Average Unit: Bytes	resource_utilization
glue.ALL.memory.heap.used.percentage	Metric Category: resource_utilization The executors' used (%) heap memory. ALL means all executors. Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (gauge), and ObservabilityGroup (resource_utilization) Valid Statistics: Average Unit: Percentage	resource_utilization

Metric	Description	Category
glue.ALL.memory.non-heap.[available | used]	Metric Category: resource_utilization The executors' available/used non-heap memory. ALL means all executors. Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (gauge), and ObservabilityGroup (resource_utilization) Valid Statistics: Average Unit: Bytes	resource_utilization
glue.ALL.memory.non-heap.used.percentage	Metric Category: resource_utilization The executors' used (%) non-heap memory. ALL means all executors. Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (gauge), and ObservabilityGroup (resource_utilization) Valid Statistics: Average Unit: Percentage	resource_utilization

Metric	Description	Category
glue.ALL.memory.total.[available | used]	Metric Category: resource_utilization The executors' available/used total memory. ALL means all executors. Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (gauge), and ObservabilityGroup (resource_utilization) Valid Statistics: Average Unit: Bytes	resource_utilization
glue.ALL.memory.total.used.percentage	Metric Category: resource_utilization The executors' used (%) total memory. ALL means all executors. Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (gauge), and ObservabilityGroup (resource_utilization) Valid Statistics: Average Unit: Percentage	resource_utilization

Metric	Description	Category
glue.driver.disk.[available_GB | used_GB]	Metric Category: resource_utilization The driver's available/used disk space during the job run. This helps to understand disk usage trends, especially over time, which can help avoid potential failures, in addition to debugging not enought disk space related failures. Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (gauge), and ObservabilityGroup (resource_utilization) Valid Statistics: Average Unit: Gigabytes	resource_utilization

Metric	Description	Category
glue.driver.disk.used.percentage]	Metric Category: resource_utilization The driver's available/used disk space during the job run. This helps to understand disk usage trends, especially over time, which can help avoid potential failures, in addition to debugging not enought disk space related failures. Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (gauge), and ObservabilityGroup (resource_utilization) Valid Statistics: Average Unit: Percentage	resource_utilization

Metric	Description	Category
glue.ALL.disk.[available_GB | used_GB]	Metric Category: resource_utilization The executors' available/used disk space. ALL means all executors. Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (gauge), and ObservabilityGroup (resource_utilization) Valid Statistics: Average Unit: Gigabytes	resource_utilization
glue.ALL.disk.used.percentage	Metric Category: resource_utilization The executors' available/used/used(%) disk space. ALL means all executors. Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (gauge), and ObservabilityGroup (resource_utilization) Valid Statistics: Average Unit: Percentage	resource_utilization

Metric	Description	Category
glue.driver.bytesRead	Metric Category: throughput The number of bytes read per input source in this job run, as well as well as for ALL sources. This helps understand the data volume and its changes over time, which helps addressing issues such as data skewness. Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (gauge), ObservabilityGroup (resource_utilization), and Source (source data location) Valid Statistics: Average Unit: Bytes	throughput

Metric	Description	Category
glue.driver.[recordsRead | filesRead]	Metric Category: throughput The number of records/files read per input source in this job run, as well as well as for ALL sources. This helps understand the data volume and its changes over time, which helps addressing issues such as data skewness. Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (gauge), ObservabilityGroup (resource _utilization), and Source (source data location) Valid Statistics: Average Unit: Count	throughput

Metric	Description	Category
glue.driver.partitionsRead	Metric Category: throughput The number of partitions read per Amazon S3 input source in this job run, as well as well as for ALL sources. Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (gauge), ObservabilityGroup (resource _utilization), and Source (source data location) Valid Statistics: Average Unit: Count	throughput

Metric	Description	Category
glue.driver.bytesWrittten	Metric Category: throughput The number of bytes written per output sink in this job run, as well as well as for ALL sinks. This helps understand the data volume and how it evolves over time, which helps addressing issues such as processing skewness. Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (gauge), ObservabilityGroup (resource_utilization), and Sink (sink data location) Valid Statistics: Average Unit: Bytes	throughput

Metric	Description	Category
glue.driver.[recordsWritten | filesWritten]	Metric Category: throughput The nnumber of records/files written per output sink in this job run, as well as well as for ALL sinks. This helps understand the data volume and how it evolves over time, which helps addressing issues such as processing skewness. Valid dimensions: JobName (the name of the AWS Glue Job), JobRunId (the JobRun ID. or ALL), Type (gauge), ObservabilityGroup (resource_utilization), and Sink (sink data location) Valid Statistics: Average Unit: Count	throughput

Error categories

Error categories	Description
COMPILATION_ERROR	Errors arise during the compilation of Scala code.
CONNECTION_ERROR	Errors arise during connecting to a service/remote host/database service, etc.
DISK_NO_SPACE_ERROR	Errors arise when there is no space left in disk on driver/executor.

Error categories	Description
OUT_OF_MEMORY_ERROR	Errors arise when there is no space left in memory on driver/executor.
IMPORT_ERROR	Errors arise when import dependencies.
INVALID_ARGUMENT_ERROR	Errors arise when the input arguments are invalid/illegal.
PERMISSION_ERROR	Errors arise when lacking the permission to service, data, etc.
RESOURCE_NOT_FOUND_ERROR	Errors arise when data, location, etc does not exit.
QUERY_ERROR	Errors arise from Spark SQL query execution.
SYNTAX_ERROR	Errors arise when there is syntax error in the script.
THROTTLING_ERROR	Errors arise when hitting service concurrency limitation or execeding service quota limitaion .
DATA_LAKE_FRAMEWORK_ERROR	Errors arise from AWS Glue native-supported data lake framework like Hudi, Iceberg, etc.
UNSUPPORTED_OPERATION_ERROR	Errors arise when making unsupported operation.
RESOURCES_ALREADY_EXISTS_ERROR	Errors arise when a resource to be created or added already exists.
GLUE_INTERNAL_SERVICE_ERROR	Errors arise when there is a AWS Glue internal service issue.
GLUE_OPERATION_TIMEOUT_ERROR	Errors arise when a AWS Glue operation is timeout.

Error categories	Description
GLUE_VALIDATION_ERROR	Errors arise when a required value could not be validated for AWS Glue job.
GLUE_JOB_BOOKMARK_VERSION_MISMATCH_ERROR	Errors arise when same job exon the same source bucket and write to the same/different destination concurrently (concurrency >1)
LAUNCH_ERROR	Errors arise during the AWS Glue job launch phase.
DYNAMODB_ERROR	Generic errors arise from Amazon DynamoDB service.
GLUE_ERROR	Generic Errors arise from AWS Glue service.
LAKEFORMATION_ERROR	Generic Errors arise from AWS Lake Formation service.
REDSHIFT_ERROR	Generic Errors arise from Amazon Redshift service.
S3_ERROR	Generic Errors arise from Amazon S3 service.
SYSTEM_EXIT_ERROR	Generic system exit error.
TIMEOUT_ERROR	Generic errors arise when job failed by operation time out.
UNCLASSIFIED_SPARK_ERROR	Generic errors arise from Spark.
UNCLASSIFIED_ERROR	Default error category.

Limitations

Note

`glueContext` must be initialized to publish the metrics.

In the Source Dimension, the value is either Amazon S3 path or table name, depending on the source type. In addition, if the source is JDBC and the query option is used, the query string is set in the source dimension. If the value is longer than 500 characters, it is trimmed within 500 characters.The following are limitations in the value:

- Non-ASCII characters will be removed.
- If the source name doesn't contain any ASCII character, it is converted to <non-ASCII input>.

Limitations and considerations for throughput metrics

- DataFrame and DataFrame-based DynamicFrame (e.g. JDBC, reading from parquet on Amazon S3) are supported, however, RDD-based DynamicFrame (e.g. reading csv, json on Amazon S3, etc.) is not supported. Technically, all reads and writes visible on Spark UI are supported.
- The `recordsRead` metric will be emitted if the data source is catalog table and the format is JSON, CSV, text, or Iceberg.
- `glue.driver.throughput.recordsWritten`, `glue.driver.throughput.bytesWritten`, and `glue.driver.throughput.filesWritten` metrics are not available in JDBC and Iceberg tables.
- Metrics may be delayed. If the job finishes in about one minute, there may be no throughput metrics in Amazon CloudWatch Metrics.

Job monitoring and debugging

You can collect metrics about AWS Glue jobs and visualize them on the AWS Glue and Amazon CloudWatch consoles to identify and fix issues. Profiling your AWS Glue jobs requires the following steps:

1. Enable metrics:
    a. Enable the **Job metrics** option in the job definition. You can enable profiling in the AWS Glue console or as a parameter to the job. For more information see Defining job properties for Spark jobs or Using job parameters in AWS Glue jobs.
    b. Enable the **AWS Glue Observability metrics** option in the job definition. You can enable Observability in the AWS Glue console or as a parameter to the job. For more information see Monitoring with AWS Glue Observability metrics.

2. Confirm that the job script initializes a `GlueContext`. For example, the following script snippet initializes a `GlueContext` and shows where profiled code is placed in the script. This general format is used in the debugging scenarios that follow.

    ```
    import sys
    from awsglue.transforms import *
    from awsglue.utils import getResolvedOptions
    from pyspark.context import SparkContext
    from awsglue.context import GlueContext
    from awsglue.job import Job
    import time

    ## @params: [JOB_NAME]
    args = getResolvedOptions(sys.argv, ['JOB_NAME'])

    sc = SparkContext()
    glueContext = GlueContext(sc)
    spark = glueContext.spark_session
    job = Job(glueContext)
    job.init(args['JOB_NAME'], args)

    ...
    ...
    code-to-profile
    ...
    ...


    job.commit()
    ```

3. Run the job.
4. Visualize the metrics:
    a. Visualize job metrics on the AWS Glue console and identify abnormal metrics for the driver or an executor.
    b. Check observability metrics in the Job run monitoring page, job run details page, or on Amazon CloudWatch. For more information, see Monitoring with AWS Glue Observability metrics.
5. Narrow down the root cause using the identified metric.
6. Optionally, confirm the root cause using the log stream of the identified driver or job executor.

Use cases for AWS Glue observability metrics

- Debugging OOM exceptions and job abnormalities
- Debugging demanding stages and straggler tasks
- Monitoring the progress of multiple jobs
- Monitoring for DPU capacity planning
- Using AWS Glue Observability for monitoring resource utilization to reduce cost

Debugging OOM exceptions and job abnormalities

You can debug out-of-memory (OOM) exceptions and job abnormalities in AWS Glue. The following sections describe scenarios for debugging out-of-memory exceptions of the Apache Spark driver or a Spark executor.

- Debugging a driver OOM exception
- Debugging an executor OOM exception

Debugging a driver OOM exception

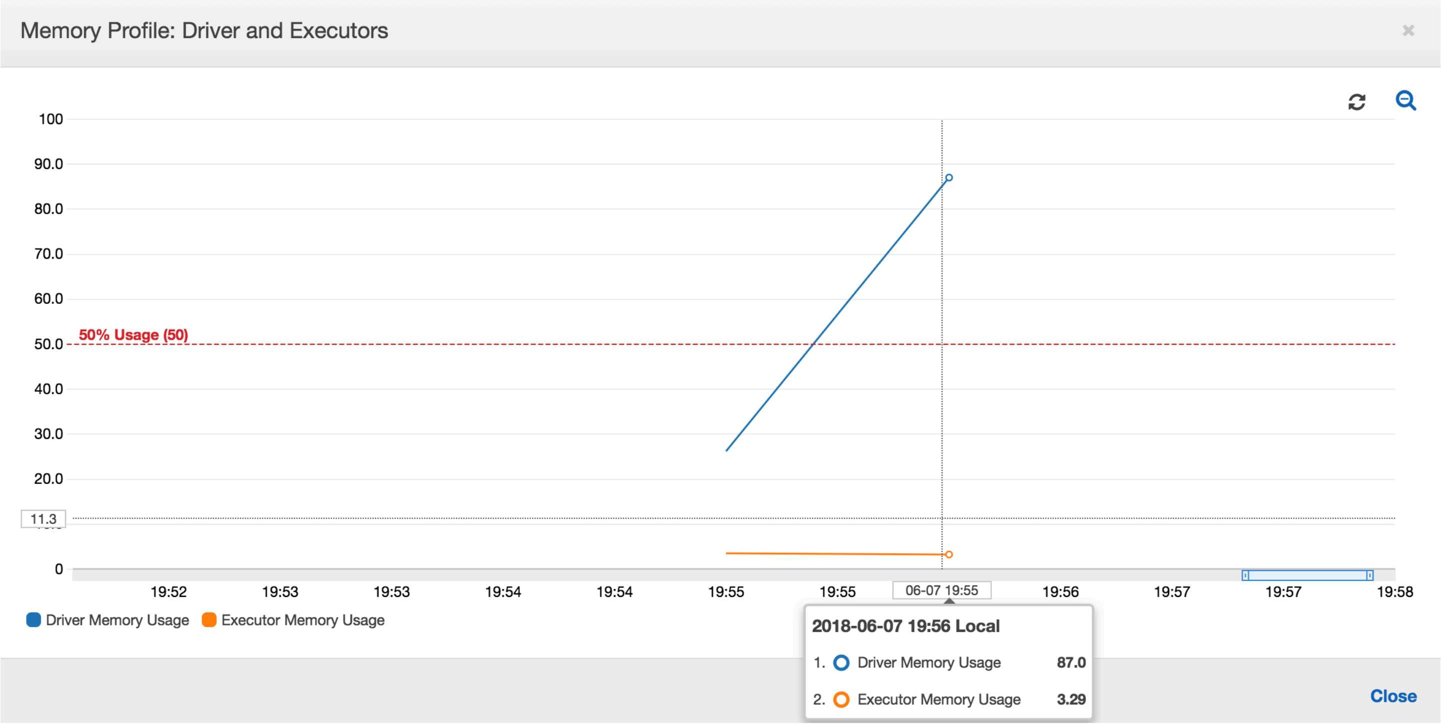
In this scenario, a Spark job is reading a large number of small files from Amazon Simple Storage Service (Amazon S3). It converts the files to Apache Parquet format and then writes them out to Amazon S3. The Spark driver is running out of memory. The input Amazon S3 data has more than 1 million files in different Amazon S3 partitions.

The profiled code is as follows:

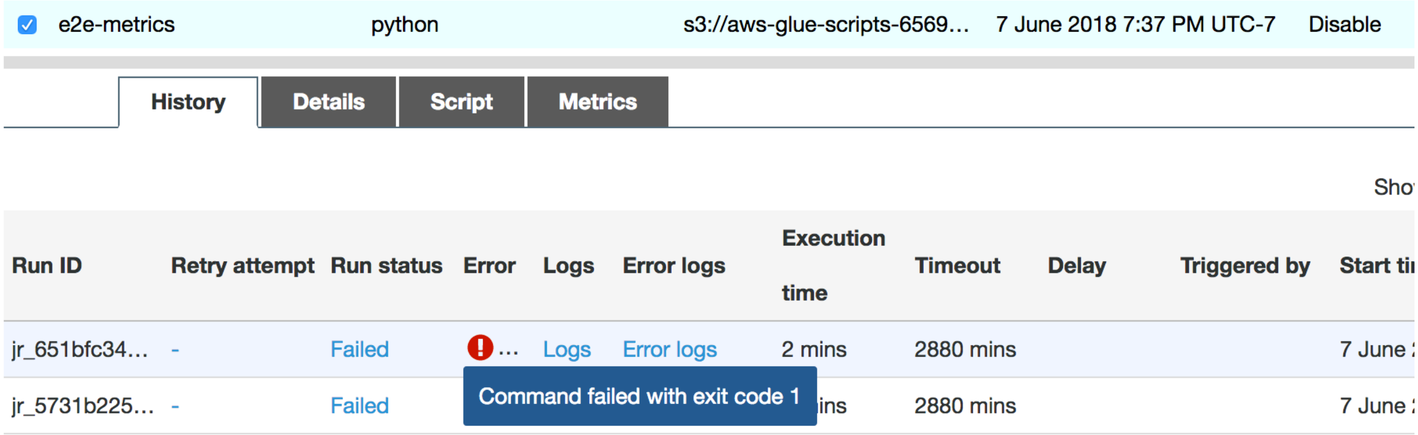
```
data = spark.read.format("json").option("inferSchema", False).load("s3://input_path")
data.write.format("parquet").save(output_path)
```

Visualize the profiled metrics on the AWS Glue console

The following graph shows the memory usage as a percentage for the driver and executors. This usage is plotted as one data point that is averaged over the values reported in the last minute. You can see in the memory profile of the job that the driver memory crosses the safe threshold of 50 percent usage quickly. On the other hand, the average memory usage across all executors is still less than 4 percent. This clearly shows abnormality with driver execution in this Spark job.



The job run soon fails, and the following error appears in the **History** tab on the AWS Glue console: Command Failed with Exit Code 1. This error string means that the job failed due to a systemic error—which in this case is the driver running out of memory.



☑ e2e-metrics | python | s3://aws-glue-scripts-6569… | 7 June 2018 7:37 PM UTC-7 | Disable

Run ID	Retry attempt	Run status	Error	Logs	Error logs	Execution time	Timeout	Delay	Triggered by	Start ti
jr_651bfc34…	-	Failed	❗…	Logs	Error logs	2 mins	2880 mins			7 June
jr_5731b225…	-	Failed				ins	2880 mins			7 June

Command failed with exit code 1

On the console, choose the **Error logs** link on the **History** tab to confirm the finding about driver OOM from the CloudWatch Logs. Search for "`Error`" in the job's error logs to confirm that it was indeed an OOM exception that failed the job:

```
# java.lang.OutOfMemoryError: Java heap space
# -XX:OnOutOfMemoryError="kill -9 %p"
# Executing /bin/sh -c "kill -9 12039"...
```

On the **History** tab for the job, choose **Logs**. You can find the following trace of driver execution in the CloudWatch Logs at the beginning of the job. The Spark driver tries to list all the files in all the directories, constructs an `InMemoryFileIndex`, and launches one task per file. This in turn results in the Spark driver having to maintain a large amount of state in memory to track all the tasks. It caches the complete list of a large number of files for the in-memory index, resulting in a driver OOM.

Fix the processing of multiple files using grouping

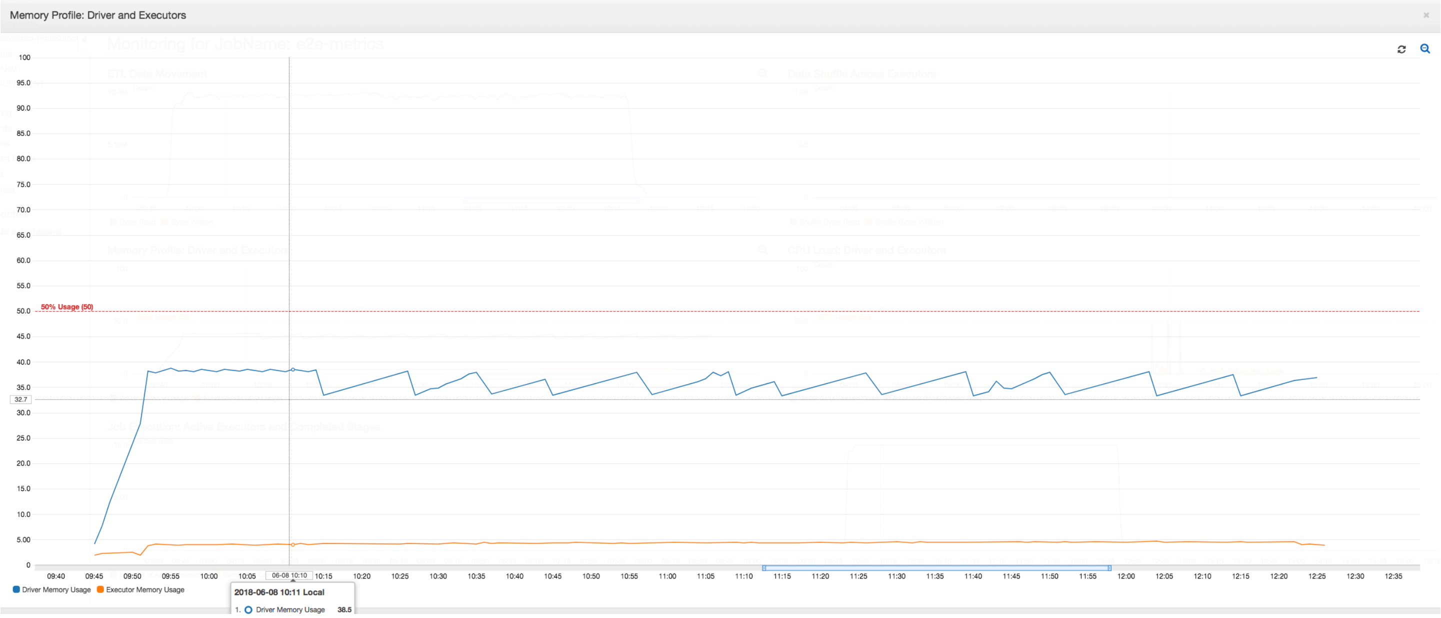
You can fix the processing of the multiple files by using the *grouping* feature in AWS Glue. Grouping is automatically enabled when you use dynamic frames and when the input dataset has a large number of files (more than 50,000). Grouping allows you to coalesce multiple files together into a group, and it allows a task to process the entire group instead of a single file. As a result, the Spark driver stores significantly less state in memory to track fewer tasks. For more information about manually enabling grouping for your dataset, see Reading input files in larger groups.

To check the memory profile of the AWS Glue job, profile the following code with grouping enabled:

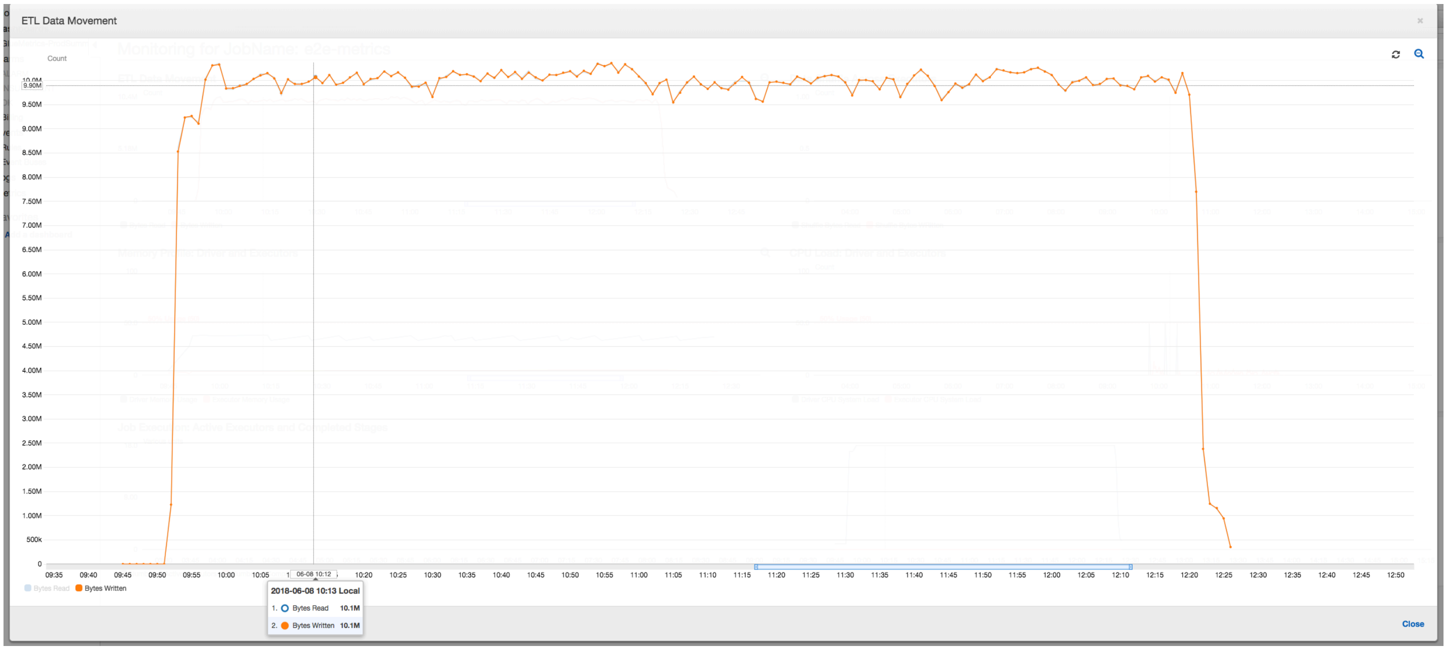
```
df = glueContext.create_dynamic_frame_from_options("s3", {'paths': ["s3://input_path"],
 "recurse":True, 'groupFiles': 'inPartition'}, format="json")
datasink = glueContext.write_dynamic_frame.from_options(frame = df, connection_type
 = "s3", connection_options = {"path": output_path}, format = "parquet",
 transformation_ctx = "datasink")
```

You can monitor the memory profile and the ETL data movement in the AWS Glue job profile.

The driver runs below the threshold of 50 percent memory usage over the entire duration of the AWS Glue job. The executors stream the data from Amazon S3, process it, and write it out to Amazon S3. As a result, they consume less than 5 percent memory at any point in time.



The data movement profile below shows the total number of Amazon S3 bytes that are read and written in the last minute by all executors as the job progresses. Both follow a similar pattern as the data is streamed across all the executors. The job finishes processing all one million files in less than three hours.



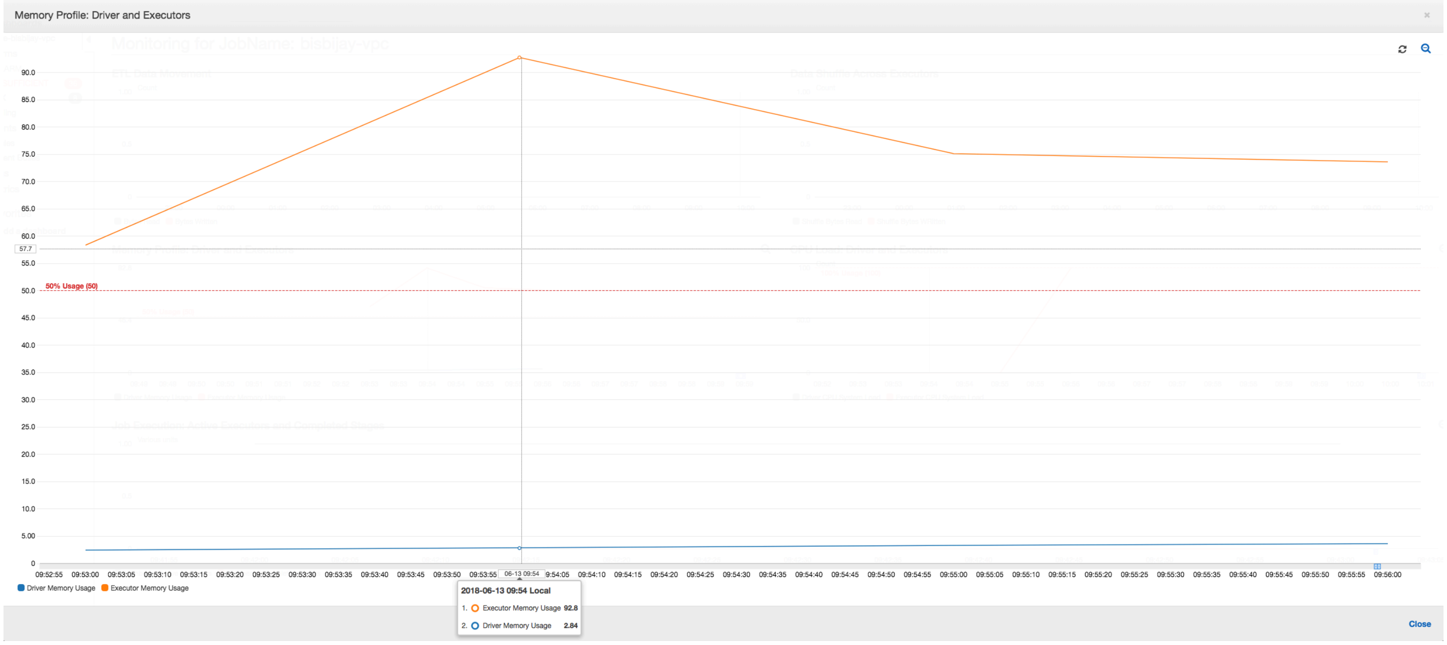
Debugging an executor OOM exception

In this scenario, you can learn how to debug OOM exceptions that could occur in Apache Spark executors. The following code uses the Spark MySQL reader to read a large table of about 34 million rows into a Spark dataframe. It then writes it out to Amazon S3 in Parquet format. You can provide the connection properties and use the default Spark configurations to read the table.

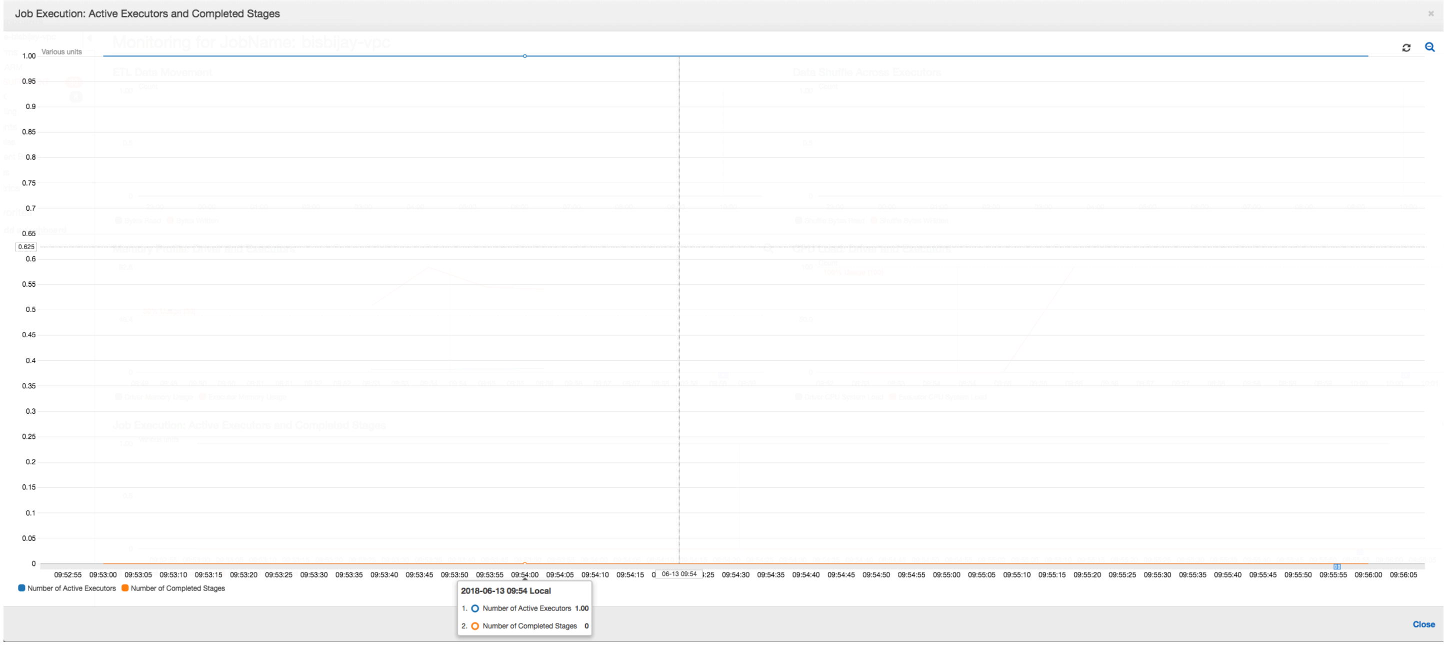
```
val connectionProperties = new Properties()
connectionProperties.put("user", user)
connectionProperties.put("password", password)
connectionProperties.put("Driver", "com.mysql.jdbc.Driver")
val sparkSession = glueContext.sparkSession
val dfSpark = sparkSession.read.jdbc(url, tableName, connectionProperties)
dfSpark.write.format("parquet").save(output_path)
```

Visualize the profiled metrics on the AWS Glue console

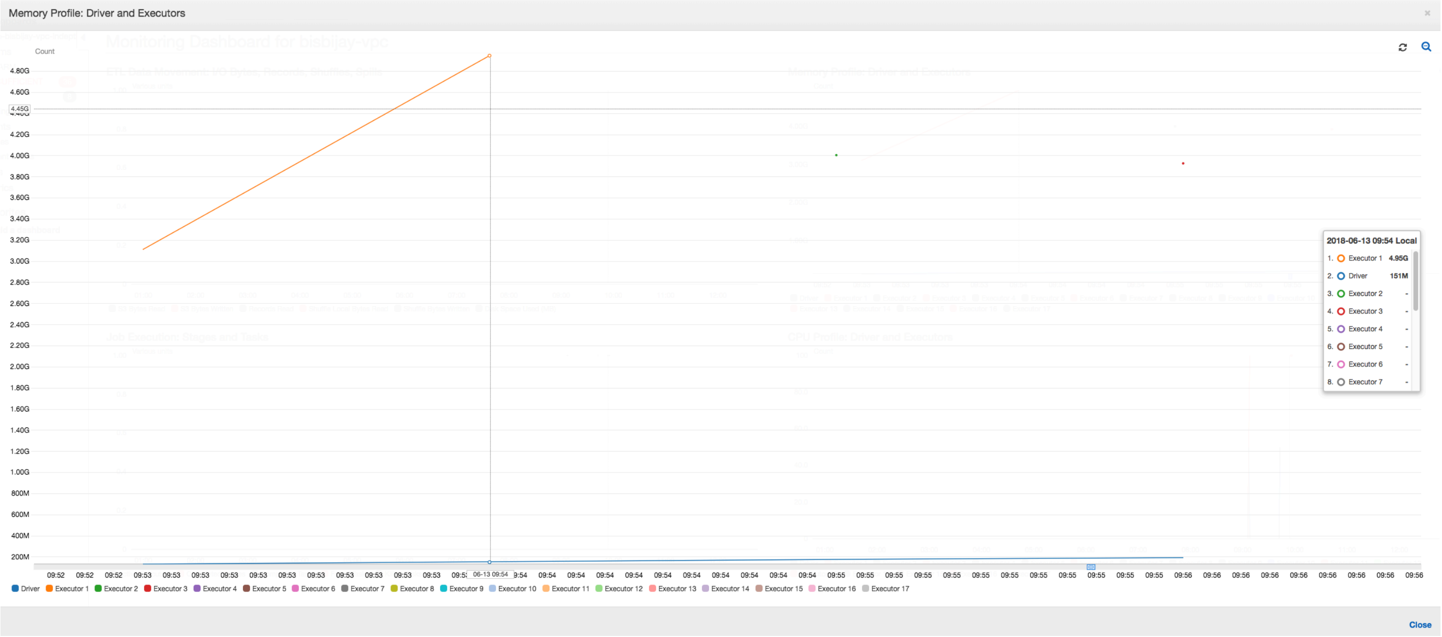
If the slope of the memory usage graph is positive and crosses 50 percent, then if the job fails before the next metric is emitted, then memory exhaustion is a good candidate for the cause. The following graph shows that within a minute of execution, the average memory usage across all executors spikes up quickly above 50 percent. The usage reaches up to 92 percent and the container running the executor is stopped by Apache Hadoop YARN.



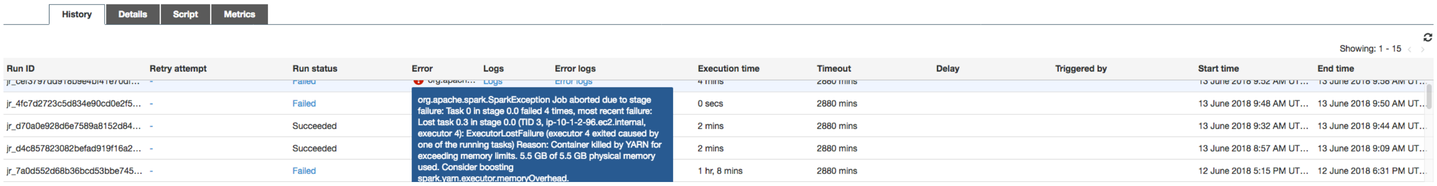
As the following graph shows, there is always a single executor running until the job fails. This is because a new executor is launched to replace the stopped executor. The JDBC data source reads are not parallelized by default because it would require partitioning the table on a column and opening multiple connections. As a result, only one executor reads in the complete table sequentially.


Job Execution: Active Executors and Completed Stages
Various units
1.00
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.625
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
09:52:55
09:56:05
Number of Active Executors
Number of Completed Stages
2018-06-13 09:54 Local
1. Number of Active Executors 1.00
2. Number of Completed Stages 0
Close

As the following graph shows, Spark tries to launch a new task four times before failing the job. You can see the memory profile of three executors. Each executor quickly uses up all of its memory. The fourth executor runs out of memory, and the job fails. As a result, its metric is not reported immediately.


Memory Profile: Driver and Executors
Count
4.80G
4.60G
4.45G
4.20G
4.00G
3.80G
3.60G
3.40G
3.20G
3.00G
2.80G
2.60G
2.40G
2.20G
2.00G
1.80G
1.60G
1.40G
1.20G
1.00G
800M
600M
400M
200M
09:52
09:56
2018-06-13 09:54 Local
1. Executor 1 4.95G
2. Driver 151M
3. Executor 2 -
4. Executor 3 -
5. Executor 4 -
6. Executor 5 -
7. Executor 6 -
8. Executor 7 -
Driver
Executor 1
Executor 2
Executor 3
Executor 4
Executor 5
Executor 6
Executor 7
Executor 8
Executor 9
Executor 10
Executor 11
Executor 12
Executor 13
Executor 14
Executor 15
Executor 16
Executor 17
Close

You can confirm from the error string on the AWS Glue console that the job failed due to OOM exceptions, as shown in the following image.



Run ID	Retry attempt	Run status	Error	Logs	Error logs	Execution time	Timeout	Delay	Triggered by	Start time	End time
[illegible]	-	Failed	org.apach…	Logs	Error logs	4 mins	2880 mins			13 June 2018 9:52 AM UT…	13 June 2018 9:58 AM UT…
jr_4fc7d2723c5d834e90cd0e2f5…	-	Failed				0 secs	2880 mins			13 June 2018 9:48 AM UT…	13 June 2018 9:50 AM UT…
jr_d70a0e928d6e7589a8152d84…	-	Succeeded				2 mins	2880 mins			13 June 2018 9:32 AM UT…	13 June 2018 9:44 AM UT…
jr_d4c857823082befad919f16a2…	-	Succeeded				2 mins	2880 mins			13 June 2018 8:57 AM UT…	13 June 2018 9:09 AM UT…
jr_7a0d552d68b36bcd53bbe745…	-	Failed				1 hr, 8 mins	2880 mins			12 June 2018 5:15 PM UT…	12 June 2018 6:31 PM UT…

org.apache.spark.SparkException Job aborted due to stage failure: Task 0 in stage 0.0 failed 4 times, most recent failure: Lost task 0.3 in stage 0.0 (TID 3, ip-10-1-2-96.ec2.internal, executor 4): ExecutorLostFailure (executor 4 exited caused by one of the running tasks) Reason: Container killed by YARN for exceeding memory limits. 5.5 GB of 5.5 GB physical memory used. Consider boosting spark.yarn.executor.memoryOverhead.

Job output logs: To further confirm your finding of an executor OOM exception, look at the CloudWatch Logs. When you search for **`Error`**, you find the four executors being stopped in roughly the same time windows as shown on the metrics dashboard. All are terminated by YARN as they exceed their memory limits.

Executor 1

```
18/06/13 16:54:29 WARN YarnAllocator: Container killed by YARN for exceeding
 memory limits. 5.5 GB of 5.5 GB physical memory used. Consider boosting
 spark.yarn.executor.memoryOverhead.
18/06/13 16:54:29 WARN YarnSchedulerBackend$YarnSchedulerEndpoint: Container killed
 by YARN for exceeding memory limits. 5.5 GB of 5.5 GB physical memory used. Consider
 boosting spark.yarn.executor.memoryOverhead.
18/06/13 16:54:29 ERROR YarnClusterScheduler: Lost executor 1 on
 ip-10-1-2-175.ec2.internal: Container killed by YARN for exceeding
 memory limits. 5.5 GB of 5.5 GB physical memory used. Consider boosting
 spark.yarn.executor.memoryOverhead.
18/06/13 16:54:29 WARN TaskSetManager: Lost task 0.0 in stage 0.0 (TID 0,
 ip-10-1-2-175.ec2.internal, executor 1): ExecutorLostFailure (executor 1
 exited caused by one of the running tasks) Reason: Container killed by YARN for
 exceeding memory limits. 5.5 GB of 5.5 GB physical memory used. Consider boosting
 spark.yarn.executor.memoryOverhead.
```

Executor 2

```
18/06/13 16:55:35 WARN YarnAllocator: Container killed by YARN for exceeding
 memory limits. 5.8 GB of 5.5 GB physical memory used. Consider boosting
 spark.yarn.executor.memoryOverhead.
18/06/13 16:55:35 WARN YarnSchedulerBackend$YarnSchedulerEndpoint: Container killed
 by YARN for exceeding memory limits. 5.8 GB of 5.5 GB physical memory used. Consider
 boosting spark.yarn.executor.memoryOverhead.
18/06/13 16:55:35 ERROR YarnClusterScheduler: Lost executor 2 on
 ip-10-1-2-16.ec2.internal: Container killed by YARN for exceeding
 memory limits. 5.8 GB of 5.5 GB physical memory used. Consider boosting
 spark.yarn.executor.memoryOverhead.
18/06/13 16:55:35 WARN TaskSetManager: Lost task 0.1 in stage 0.0 (TID 1,
 ip-10-1-2-16.ec2.internal, executor 2): ExecutorLostFailure (executor 2 exited
 caused by one of the running tasks) Reason: Container killed by YARN for
 exceeding memory limits. 5.8 GB of 5.5 GB physical memory used. Consider boosting
 spark.yarn.executor.memoryOverhead.
```

Executor 3

```
18/06/13 16:56:37 WARN YarnAllocator: Container killed by YARN for exceeding
 memory limits. 5.8 GB of 5.5 GB physical memory used. Consider boosting
 spark.yarn.executor.memoryOverhead.
18/06/13 16:56:37 WARN YarnSchedulerBackend$YarnSchedulerEndpoint: Container killed
 by YARN for exceeding memory limits. 5.8 GB of 5.5 GB physical memory used. Consider
 boosting spark.yarn.executor.memoryOverhead.
18/06/13 16:56:37 ERROR YarnClusterScheduler: Lost executor 3 on
 ip-10-1-2-189.ec2.internal: Container killed by YARN for exceeding
 memory limits. 5.8 GB of 5.5 GB physical memory used. Consider boosting
 spark.yarn.executor.memoryOverhead.
18/06/13 16:56:37 WARN TaskSetManager: Lost task 0.2 in stage 0.0 (TID 2,
 ip-10-1-2-189.ec2.internal, executor 3): ExecutorLostFailure (executor 3
 exited caused by one of the running tasks) Reason: Container killed by YARN for
 exceeding memory limits. 5.8 GB of 5.5 GB physical memory used. Consider boosting
 spark.yarn.executor.memoryOverhead.
```

Executor 4

```
18/06/13 16:57:18 WARN YarnAllocator: Container killed by YARN for exceeding
 memory limits. 5.5 GB of 5.5 GB physical memory used. Consider boosting
 spark.yarn.executor.memoryOverhead.
18/06/13 16:57:18 WARN YarnSchedulerBackend$YarnSchedulerEndpoint: Container killed
 by YARN for exceeding memory limits. 5.5 GB of 5.5 GB physical memory used. Consider
 boosting spark.yarn.executor.memoryOverhead.
18/06/13 16:57:18 ERROR YarnClusterScheduler: Lost executor 4 on
 ip-10-1-2-96.ec2.internal: Container killed by YARN for exceeding
 memory limits. 5.5 GB of 5.5 GB physical memory used. Consider boosting
 spark.yarn.executor.memoryOverhead.
18/06/13 16:57:18 WARN TaskSetManager: Lost task 0.3 in stage 0.0 (TID 3,
 ip-10-1-2-96.ec2.internal, executor 4): ExecutorLostFailure (executor 4 exited
 caused by one of the running tasks) Reason: Container killed by YARN for
 exceeding memory limits. 5.5 GB of 5.5 GB physical memory used. Consider boosting
 spark.yarn.executor.memoryOverhead.
```

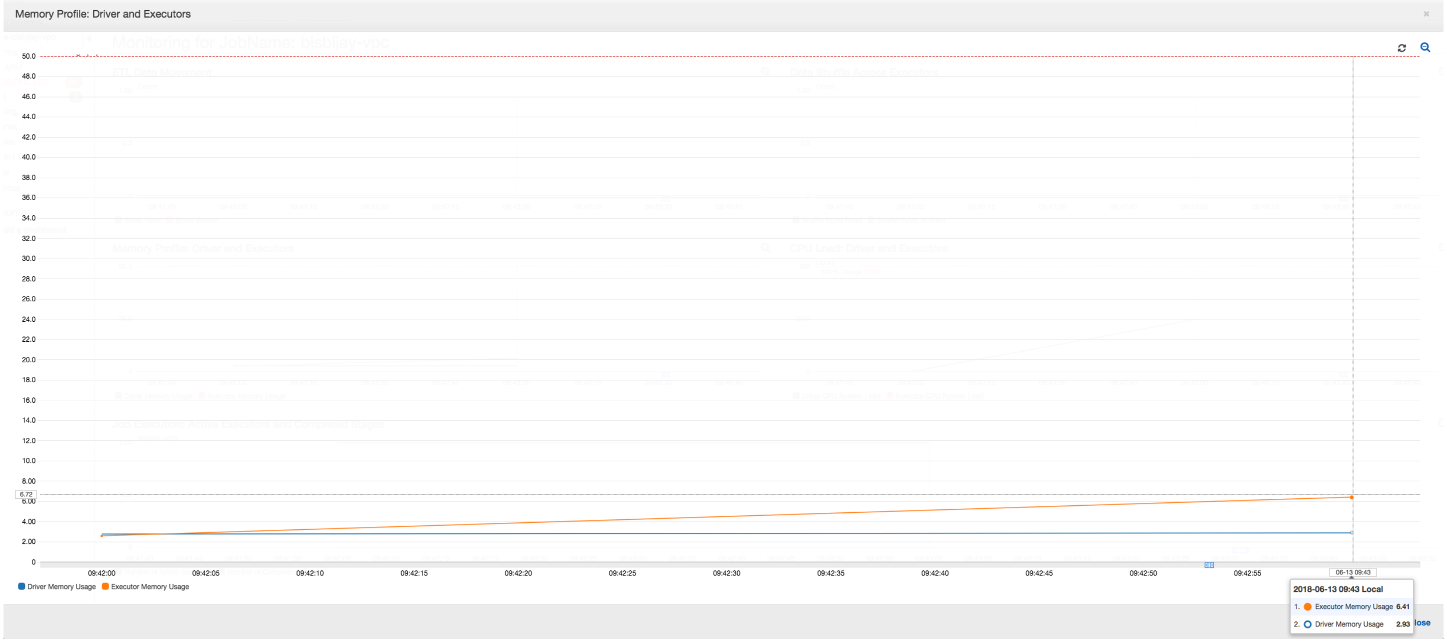
Fix the fetch size setting using AWS Glue dynamic frames

The executor ran out of memory while reading the JDBC table because the default configuration for the Spark JDBC fetch size is zero. This means that the JDBC driver on the Spark executor tries to fetch the 34 million rows from the database together and cache them, even though Spark streams through the rows one at a time. With Spark, you can avoid this scenario by setting the fetch size parameter to a non-zero default value.

You can also fix this issue by using AWS Glue dynamic frames instead. By default, dynamic frames use a fetch size of 1,000 rows that is a typically sufficient value. As a result, the executor does not take more than 7 percent of its total memory. The AWS Glue job finishes in less than two minutes with only a single executor. While using AWS Glue dynamic frames is the recommended approach, it is also possible to set the fetch size using the Apache Spark `fetchsize` property. See the Spark SQL, DataFrames and Datasets Guide.

```
val (url, database, tableName) = {
 ("jdbc_url", "db_name", "table_name")
 }
val source = glueContext.getSource(format, sourceJson)
val df = source.getDynamicFrame
glueContext.write_dynamic_frame.from_options(frame = df, connection_type = "s3",
 connection_options = {"path": output_path}, format = "parquet", transformation_ctx =
 "datasink")
```

Normal profiled metrics: The executor memory with AWS Glue dynamic frames never exceeds the safe threshold, as shown in the following image. It streams in the rows from the database and caches only 1,000 rows in the JDBC driver at any point in time. An out of memory exception does not occur.



Debugging demanding stages and straggler tasks

You can use AWS Glue job profiling to identify demanding stages and straggler tasks in your extract, transform, and load (ETL) jobs. A straggler task takes much longer than the rest of the tasks in a stage of an AWS Glue job. As a result, the stage takes longer to complete, which also delays the total execution time of the job.

Coalescing small input files into larger output files

A straggler task can occur when there is a non-uniform distribution of work across the different tasks, or a data skew results in one task processing more data.

You can profile the following code—a common pattern in Apache Spark—to coalesce a large number of small files into larger output files. For this example, the input dataset is 32 GB of JSON Gzip compressed files. The output dataset has roughly 190 GB of uncompressed JSON files.

The profiled code is as follows:

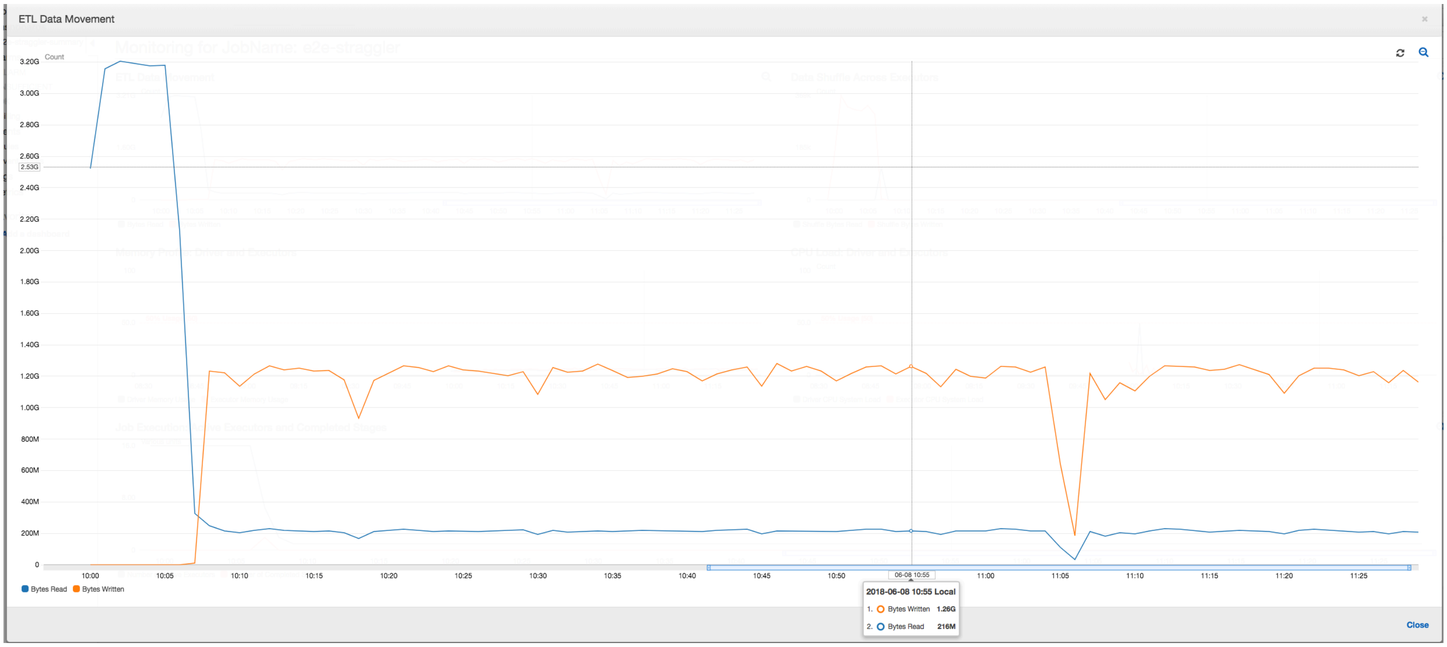
```
datasource0 = spark.read.format("json").load("s3://input_path")
df = datasource0.coalesce(1)
df.write.format("json").save(output_path)
```

Visualize the profiled metrics on the AWS Glue console

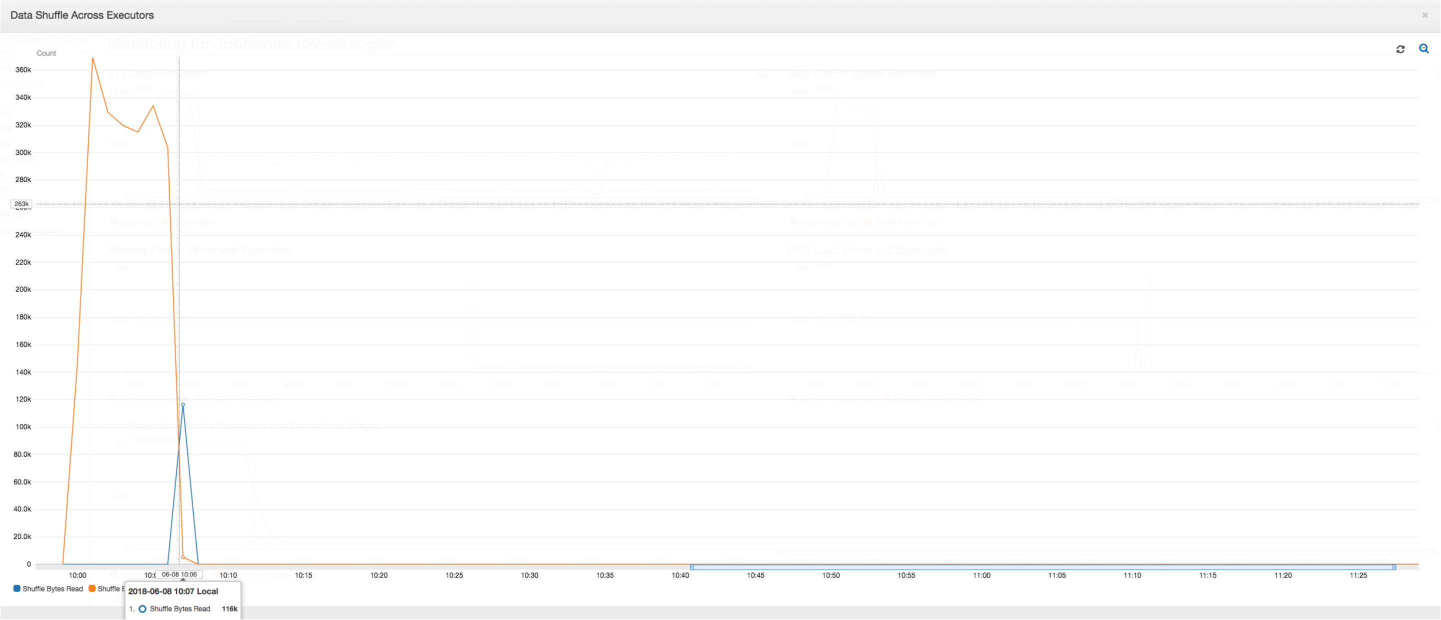
You can profile your job to examine four different sets of metrics:

- ETL data movement
- Data shuffle across executors
- Job execution
- Memory profile

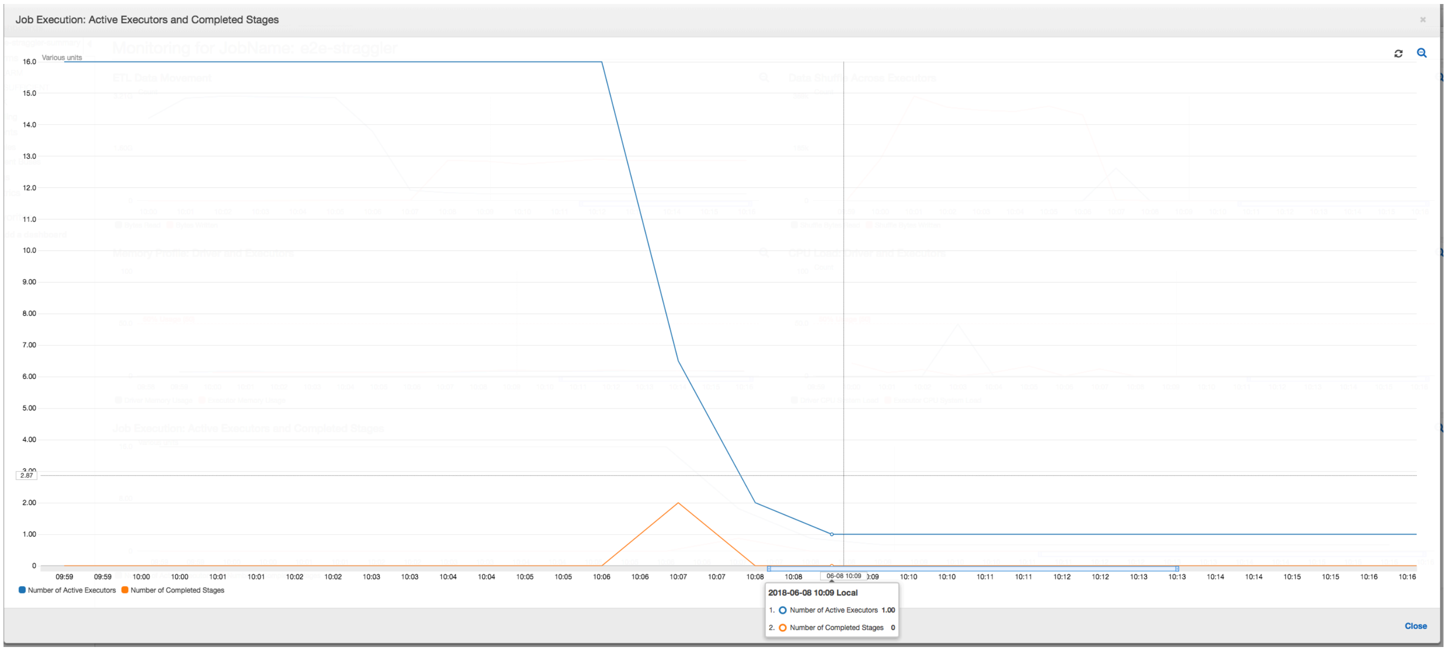
ETL data movement: In the **ETL Data Movement** profile, the bytes are read fairly quickly by all the executors in the first stage that completes within the first six minutes. However, the total job execution time is around one hour, mostly consisting of the data writes.



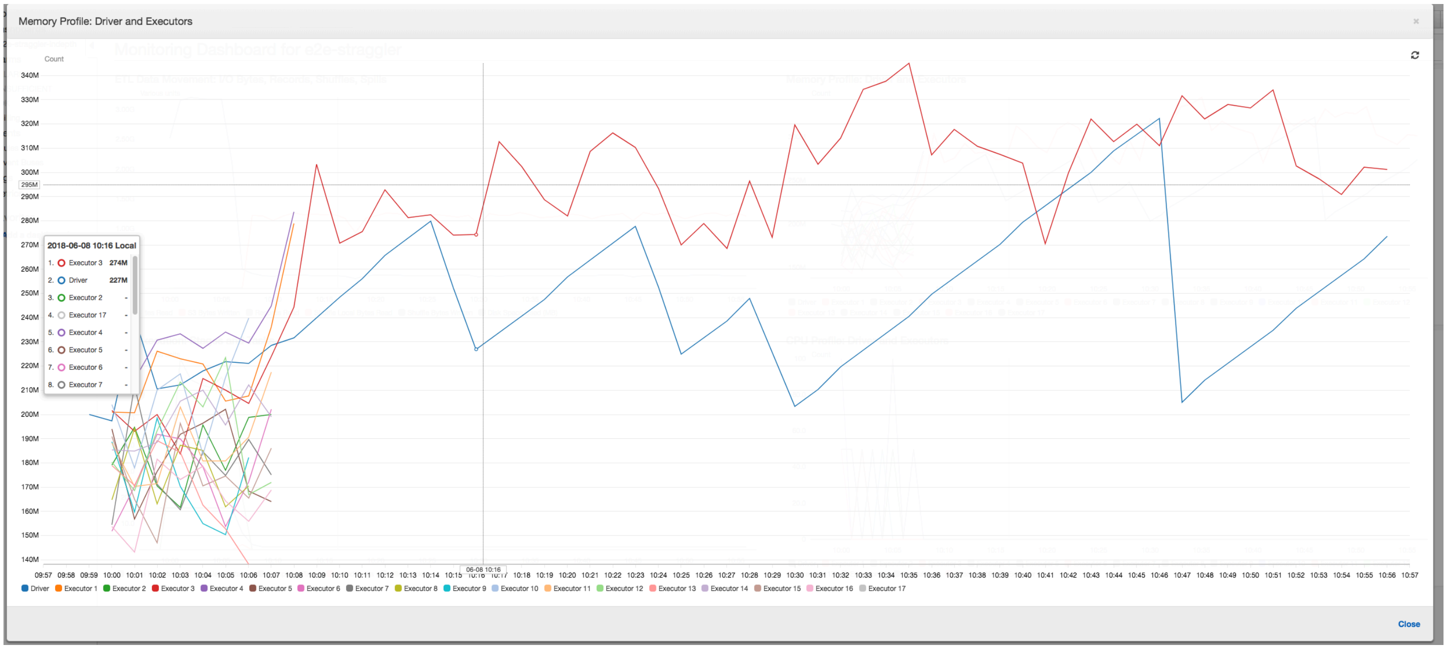
Data shuffle across executors: The number of bytes read and written during shuffling also shows a spike before Stage 2 ends, as indicated by the **Job Execution** and **Data Shuffle** metrics. After the data shuffles from all executors, the reads and writes proceed from executor number 3 only.



Job execution: As shown in the graph below, all other executors are idle and are eventually relinquished by the time 10:09. At that point, the total number of executors decreases to only one. This clearly shows that executor number 3 consists of the straggler task that is taking the longest execution time and is contributing to most of the job execution time.



Memory profile: After the first two stages, only executor number 3 is actively consuming memory to process the data. The remaining executors are simply idle or have been relinquished shortly after the completion of the first two stages.



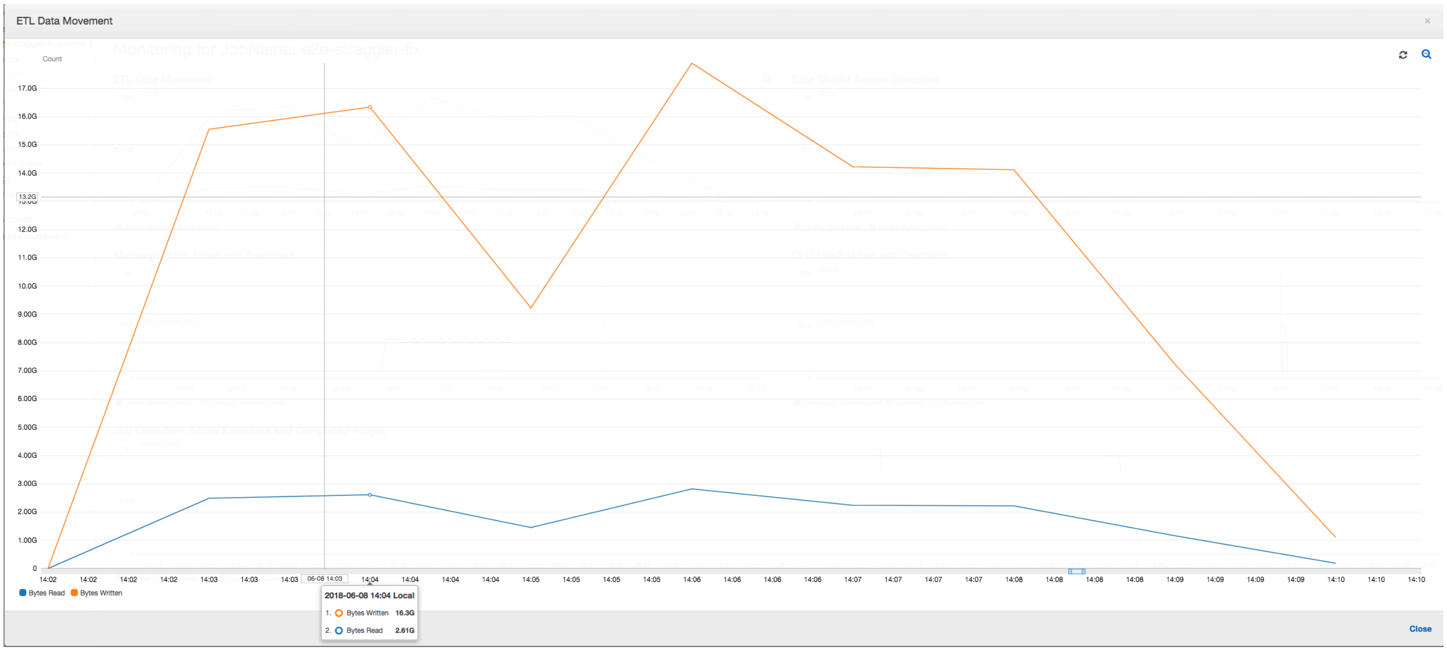
Fix straggling executors using grouping

You can avoid straggling executors by using the *grouping* feature in AWS Glue. Use grouping to distribute the data uniformly across all the executors and coalesce files into larger files using all the available executors on the cluster. For more information, see Reading input files in larger groups.

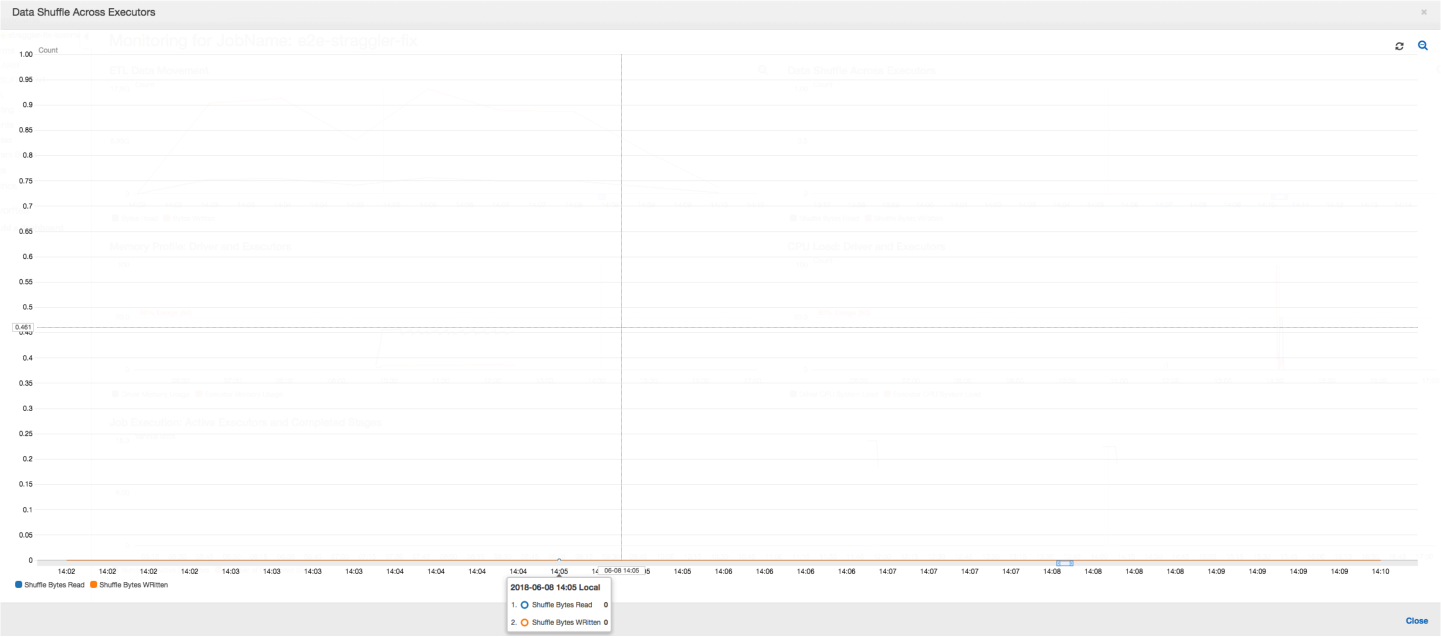
To check the ETL data movements in the AWS Glue job, profile the following code with grouping enabled:

```
df = glueContext.create_dynamic_frame_from_options("s3", {'paths': ["s3://input_path"],
 "recurse":True, 'groupFiles': 'inPartition'}, format="json")
datasink = glueContext.write_dynamic_frame.from_options(frame = df, connection_type =
 "s3", connection_options = {"path": output_path}, format = "json", transformation_ctx
 = "datasink4")
```

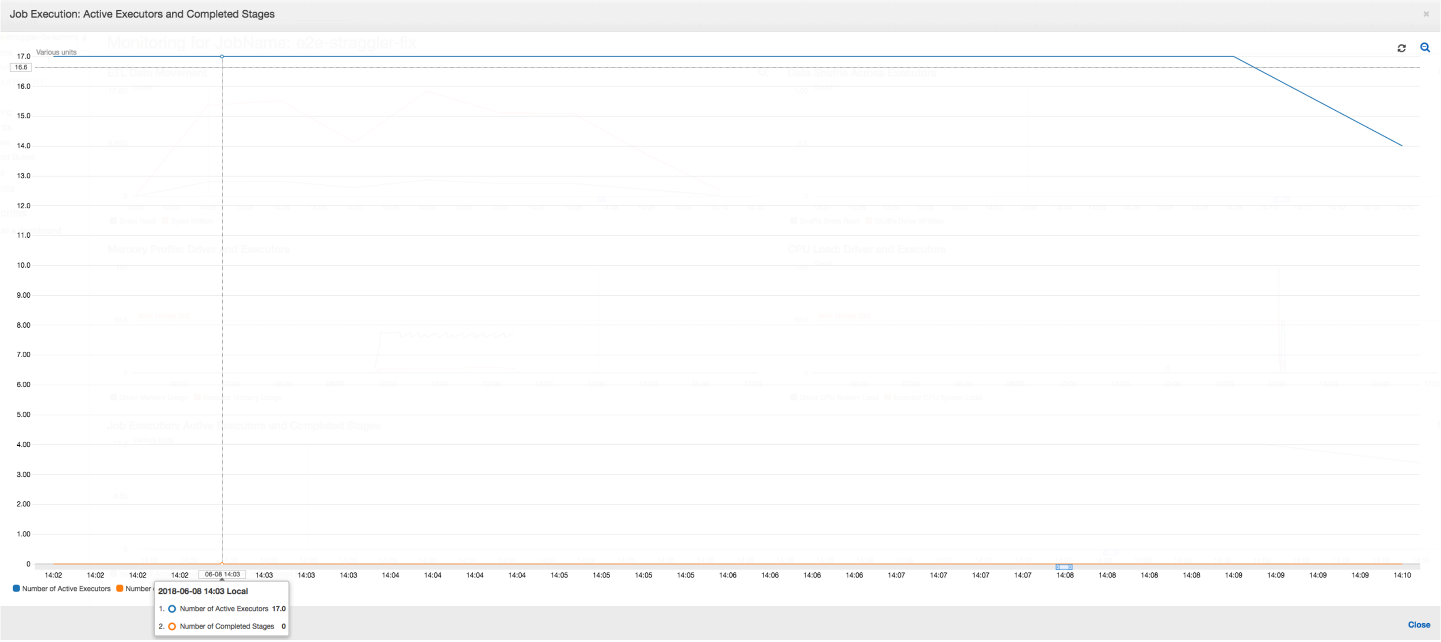
ETL data movement: The data writes are now streamed in parallel with the data reads throughout the job execution time. As a result, the job finishes within eight minutes—much faster than previously.



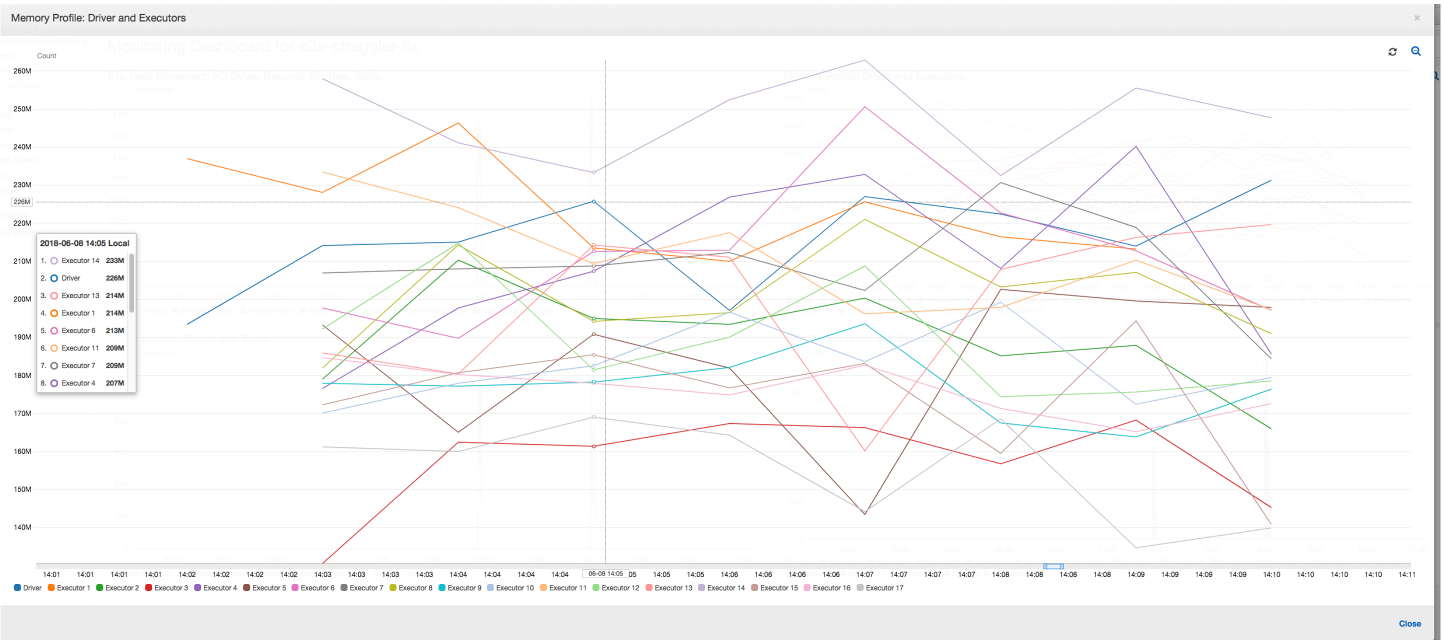
Data shuffle across executors: As the input files are coalesced during the reads using the grouping feature, there is no costly data shuffle after the data reads.



Job execution: The job execution metrics show that the total number of active executors running and processing data remains fairly constant. There is no single straggler in the job. All executors are active and are not relinquished until the completion of the job. Because there is no intermediate shuffle of data across the executors, there is only a single stage in the job.



Memory profile: The metrics show the active memory consumption across all executors—reconfirming that there is activity across all executors. As data is streamed in and written out in parallel, the total memory footprint of all executors is roughly uniform and well below the safe threshold for all executors.



Monitoring the progress of multiple jobs

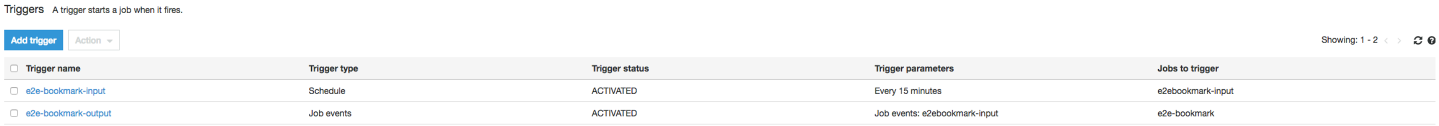
You can profile multiple AWS Glue jobs together and monitor the flow of data between them. This is a common workflow pattern, and requires monitoring for individual job progress, data processing backlog, data reprocessing, and job bookmarks.

Topics

- Profiled code
- Visualize the profiled metrics on the AWS Glue console
- Fix the processing of files

Profiled code

In this workflow, you have two jobs: an Input job and an Output job. The Input job is scheduled to run every 30 minutes using a periodic trigger. The Output job is scheduled to run after each successful run of the Input job. These scheduled jobs are controlled using job triggers.



☐ Trigger name	Trigger type	Trigger status	Trigger parameters	Jobs to trigger
☐ e2e-bookmark-input	Schedule	ACTIVATED	Every 15 minutes	e2ebookmark-input
☐ e2e-bookmark-output	Job events	ACTIVATED	Job events: e2ebookmark-input	e2e-bookmark

Input job: This job reads in data from an Amazon Simple Storage Service (Amazon S3) location, transforms it using `ApplyMapping`, and writes it to a staging Amazon S3 location. The following code is profiled code for the Input job:

```
datasource0 = glueContext.create_dynamic_frame.from_options(connection_type="s3",
 connection_options = {"paths": ["s3://input_path"],
 "useS3ListImplementation":True,"recurse":True}, format="json")
applymapping1 = ApplyMapping.apply(frame = datasource0, mappings = [map_spec])
datasink2 = glueContext.write_dynamic_frame.from_options(frame = applymapping1,
 connection_type = "s3", connection_options = {"path": staging_path, "compression":
 "gzip"}, format = "json")
```

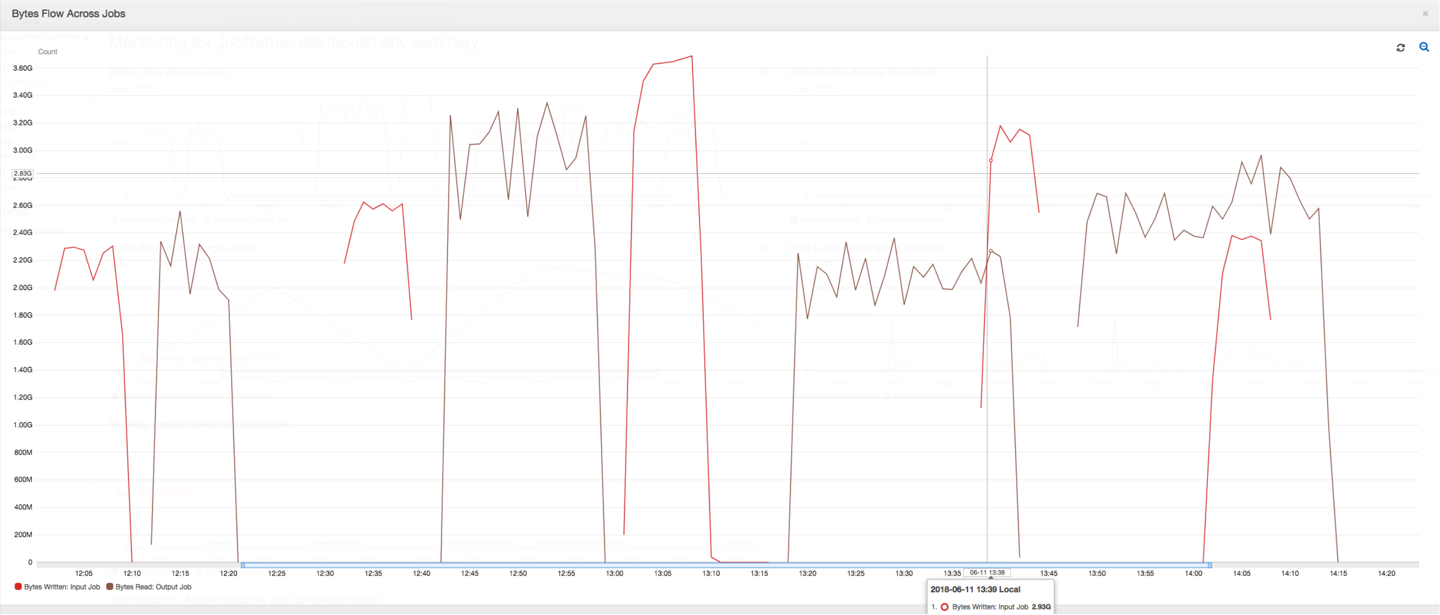
Output job: This job reads the output of the Input job from the staging location in Amazon S3, transforms it again, and writes it to a destination:

```

datasource0 = glueContext.create_dynamic_frame.from_options(connection_type="s3",
 connection_options = {"paths": [staging_path],
 "useS3ListImplementation":True,"recurse":True}, format="json")
```

```
applymapping1 = ApplyMapping.apply(frame = datasource0, mappings = [map_spec])
datasink2 = glueContext.write_dynamic_frame.from_options(frame = applymapping1,
 connection_type = "s3", connection_options = {"path": output_path}, format = "json")
```

Visualize the profiled metrics on the AWS Glue console

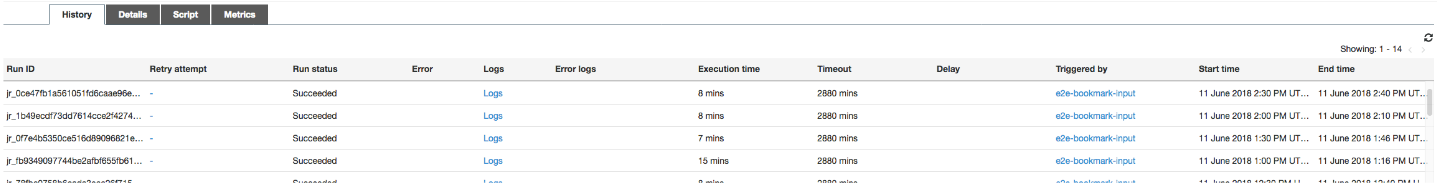
The following dashboard superimposes the Amazon S3 bytes written metric from the Input job onto the Amazon S3 bytes read metric on the same timeline for the Output job. The timeline shows different job runs of the Input and Output jobs. The Input job (shown in red) starts every 30 minutes. The Output Job (shown in brown) starts at the completion of the Input Job, with a Max Concurrency of 1.



In this example, job bookmarks are not enabled. No transformation contexts are used to enable job bookmarks in the script code.

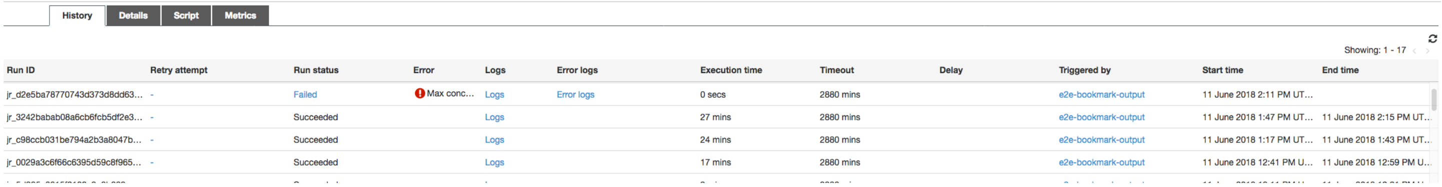
Job History: The Input and Output jobs have multiple runs, as shown on the **History** tab, starting from 12:00 PM.

The Input job on the AWS Glue console looks like this:



Run ID	Retry attempt	Run status	Error	Logs	Error logs	Execution time	Timeout	Delay	Triggered by	Start time	End time
jr_0ce47fb1a561051fd6caae96e...	-	Succeeded		Logs		8 mins	2880 mins		e2e-bookmark-input	11 June 2018 2:30 PM UT...	11 June 2018 2:40 PM UT...
jr_1b49ecdf73dd7614cce2f4274...	-	Succeeded		Logs		8 mins	2880 mins		e2e-bookmark-input	11 June 2018 2:00 PM UT...	11 June 2018 2:10 PM UT...
jr_0f7e4b5350ce516d89096821e...	-	Succeeded		Logs		7 mins	2880 mins		e2e-bookmark-input	11 June 2018 1:30 PM UT...	11 June 2018 1:46 PM UT...
jr_fb9349097744be2afbf655fb61...	-	Succeeded		Logs		15 mins	2880 mins		e2e-bookmark-input	11 June 2018 1:00 PM UT...	11 June 2018 1:16 PM UT...

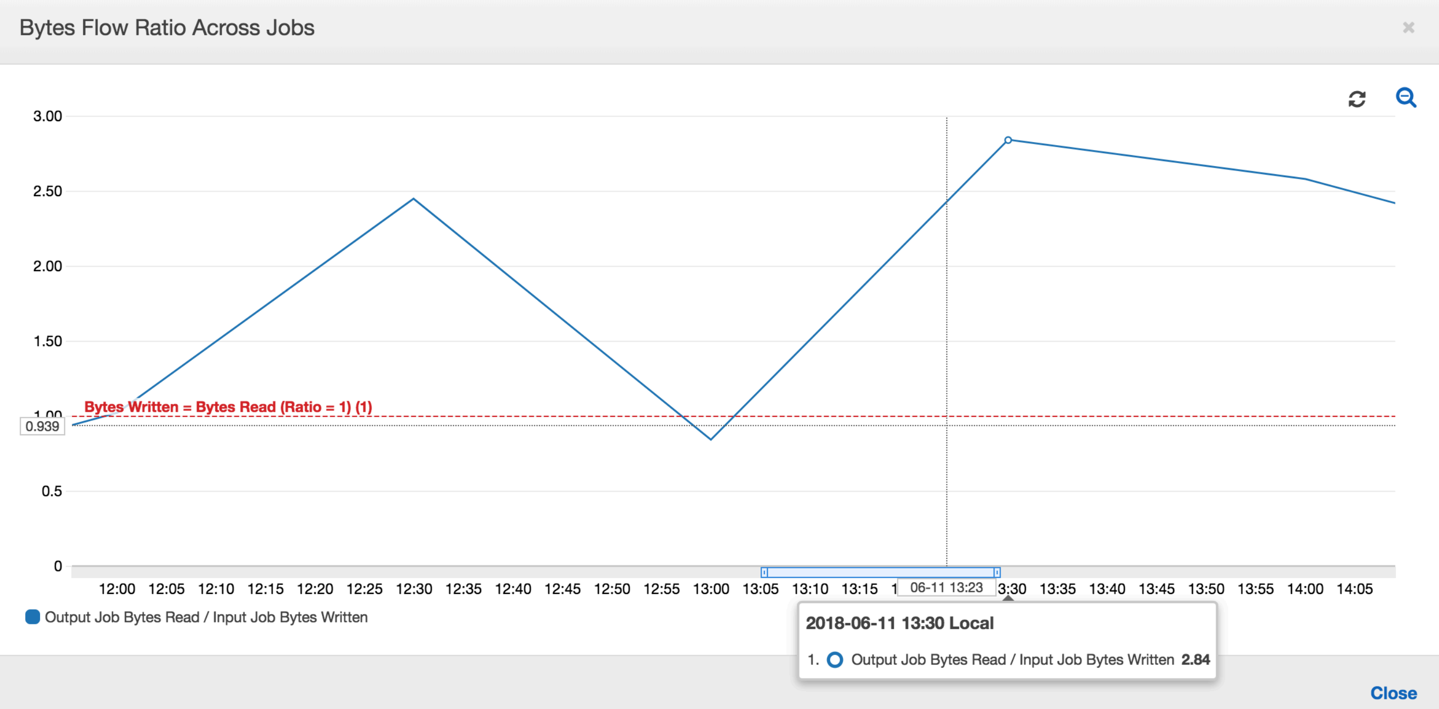
The following image shows the Output job:



Run ID	Retry attempt	Run status	Error	Logs	Error logs	Execution time	Timeout	Delay	Triggered by	Start time	End time
jr_d2e5ba78770743d373d8dd63...	-	Failed	Max conc...	Logs	Error logs	0 secs	2880 mins		e2e-bookmark-output	11 June 2018 2:11 PM UT...	
jr_3242babab08a6cb6fcb5df2e3...	-	Succeeded		Logs		27 mins	2880 mins		e2e-bookmark-output	11 June 2018 1:47 PM UT...	11 June 2018 2:15 PM UT...
jr_c98ccb031be794a2b3a8047b...	-	Succeeded		Logs		24 mins	2880 mins		e2e-bookmark-output	11 June 2018 1:17 PM UT...	11 June 2018 1:43 PM UT...
jr_0029a3c6f66c6395d59c8f965...	-	Succeeded		Logs		17 mins	2880 mins		e2e-bookmark-output	11 June 2018 12:41 PM U...	11 June 2018 12:59 PM U...

First job runs: As shown in the Data Bytes Read and Written graph below, the first job runs of the Input and Output jobs between 12:00 and 12:30 show roughly the same area under the curves. Those areas represent the Amazon S3 bytes written by the Input job and the Amazon S3 bytes read by the Output job. This data is also confirmed by the ratio of Amazon S3 bytes written (summed over 30 minutes – the job trigger frequency for the Input job). The data point for the ratio for the Input job run that started at 12:00PM is also 1.

The following graph shows the data flow ratio across all the job runs:



Second job runs: In the second job run, there is a clear difference in the number of bytes read by the Output job compared to the number of bytes written by the Input job. (Compare the area under the curve across the two job runs for the Output job, or compare the areas in the second run of the Input and Output jobs.) The ratio of the bytes read and written shows that the Output Job read about 2.5x the data written by the Input job in the second span of 30 minutes from 12:30 to 13:00. This is because the Output Job reprocessed the output of the first job run of the Input job because job bookmarks were not enabled. A ratio above 1 shows that there is an additional backlog of data that was processed by the Output job.

Third job runs: The Input job is fairly consistent in terms of the number of bytes written (see the area under the red curves). However, the third job run of the Input job ran longer than expected (see the long tail of the red curve). As a result, the third job run of the Output job started late. The third job run processed only a fraction of the data accumulated in the staging location in the remaining 30 minutes between 13:00 and 13:30. The ratio of the bytes flow shows that it only processed 0.83 of data written by the third job run of the Input job (see the ratio at 13:00).

Overlap of Input and Output jobs: The fourth job run of the Input job started at 13:30 as per the schedule, before the third job run of the Output job finished. There is a partial overlap between these two job runs. However, the third job run of the Output job captures only the files that it listed in the staging location of Amazon S3 when it began around 13:17. This consists of all data output from the first job runs of the Input job. The actual ratio at 13:30 is around 2.75. The third job run of the Output job processed about 2.75x of data written by the fourth job run of the Input job from 13:30 to 14:00.

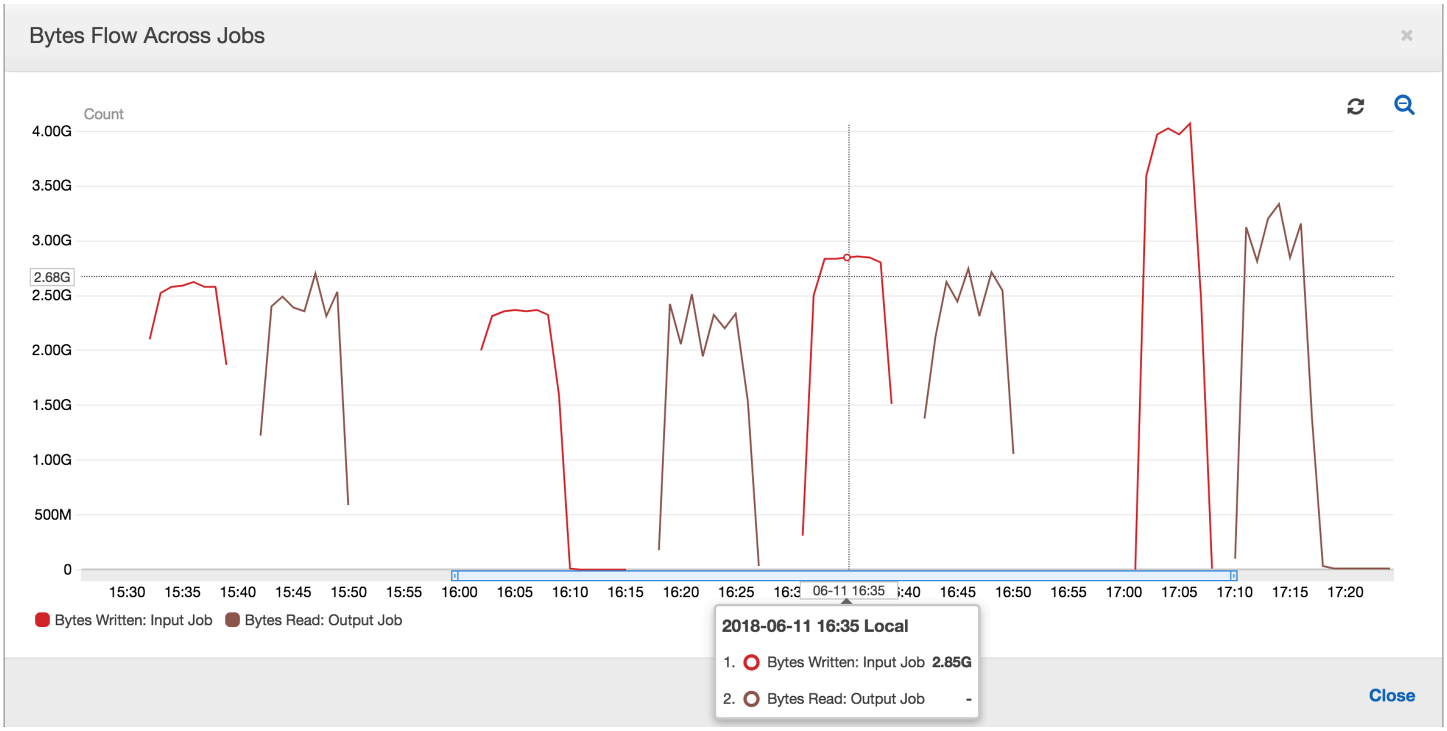
As these images show, the Output job is reprocessing data from the staging location from all prior job runs of the Input job. As a result, the fourth job run for the Output job is the longest and overlaps with the entire fifth job run of the Input job.

Fix the processing of files

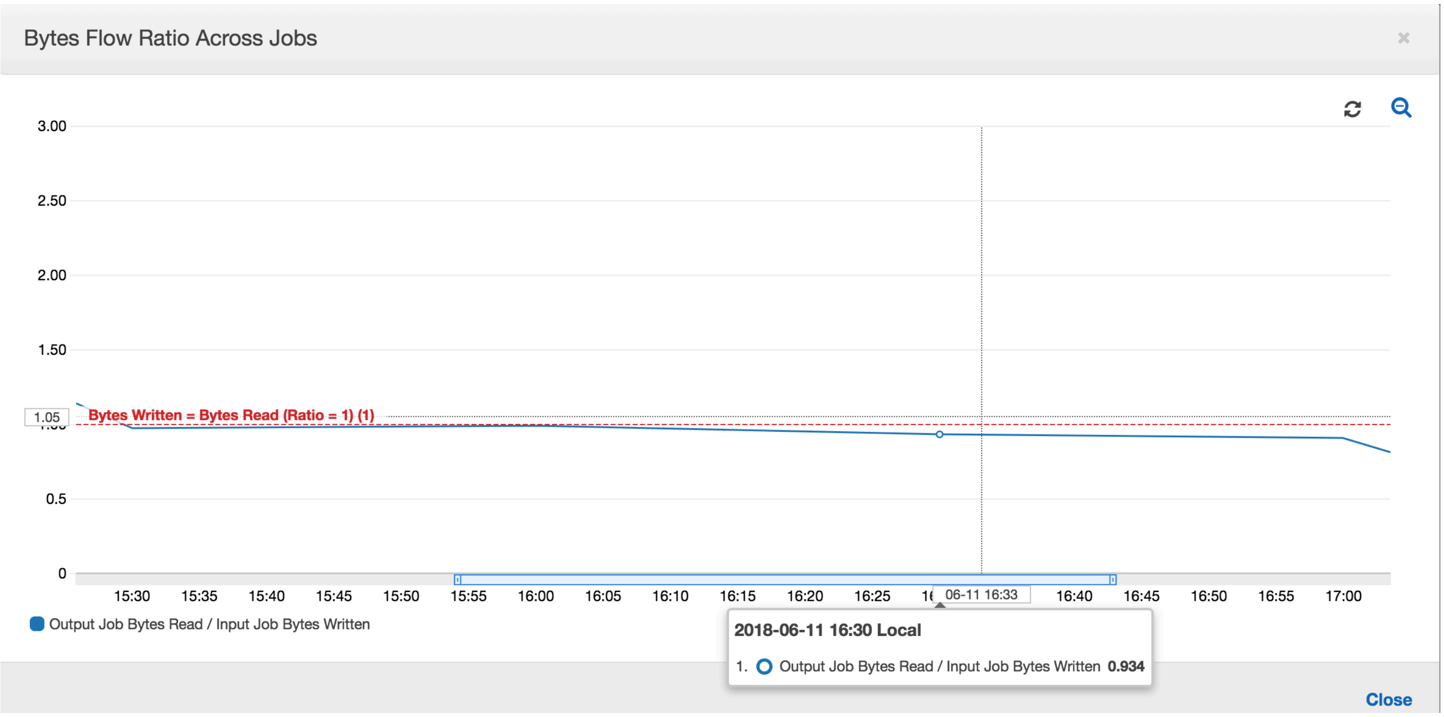
You should ensure that the Output job processes only the files that haven't been processed by previous job runs of the Output job. To do this, enable job bookmarks and set the transformation context in the Output job, as follows:

```
datasource0 = glueContext.create_dynamic_frame.from_options(connection_type="s3",
 connection_options = {"paths": [staging_path],
 "useS3ListImplementation":True,"recurse":True}, format="json", transformation_ctx =
 "bookmark_ctx")
```

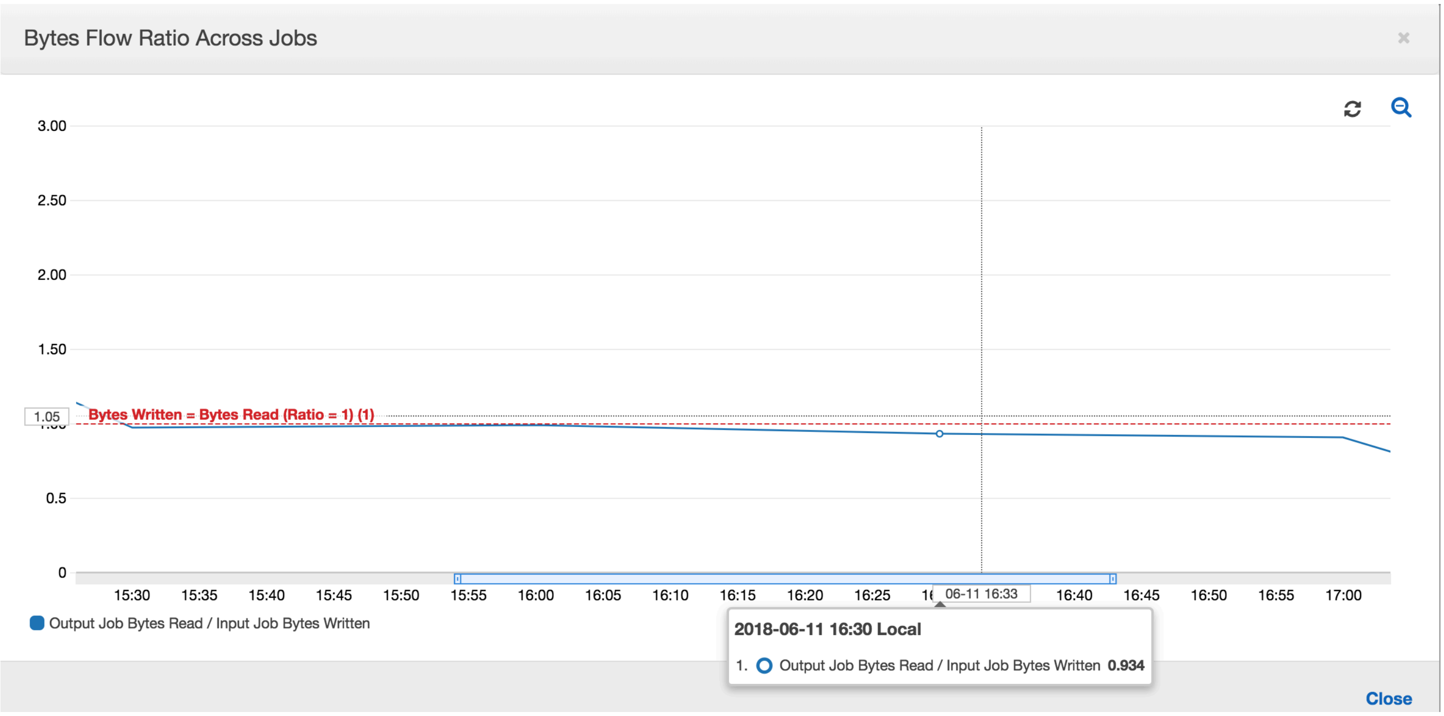
With job bookmarks enabled, the Output job doesn't reprocess the data in the staging location from all the previous job runs of the Input job. In the following image showing the data read and written, the area under the brown curve is fairly consistent and similar with the red curves.



The ratios of byte flow also remain roughly close to 1 because there is no additional data processed.



A job run for the Output job starts and captures the files in the staging location before the next Input job run starts putting more data into the staging location. As long as it continues to do this, it processes only the files captured from the previous Input job run, and the ratio stays close to 1.


Bytes Flow Ratio Across Jobs
3.00
2.50
2.00
1.50
1.05
Bytes Written = Bytes Read (Ratio = 1) (1)
0.5
0
15:30
15:35
15:40
15:45
15:50
15:55
16:00
16:05
16:10
16:15
16:20
16:25
06-11 16:33
16:40
16:45
16:50
16:55
17:00
Output Job Bytes Read / Input Job Bytes Written
2018-06-11 16:30 Local
1. Output Job Bytes Read / Input Job Bytes Written 0.934
Close

Suppose that the Input job takes longer than expected, and as a result, the Output job captures files in the staging location from two Input job runs. The ratio is then higher than 1 for that Output job run. However, the following job runs of the Output job don't process any files that are already processed by the previous job runs of the Output job.

Monitoring for DPU capacity planning

You can use job metrics in AWS Glue to estimate the number of data processing units (DPUs) that can be used to scale out an AWS Glue job.

> **Note**
>
> This page is only applicable to AWS Glue versions 0.9 and 1.0. Later versions of AWS Glue contain cost-saving features that introduce additional considerations when capacity planning.

Topics

- Profiled code
- Visualize the profiled metrics on the AWS Glue console
- Determine the optimal DPU capacity

Profiled code

The following script reads an Amazon Simple Storage Service (Amazon S3) partition containing 428 gzipped JSON files. The script applies a mapping to change the field names, and converts and writes them to Amazon S3 in Apache Parquet format. You provision 10 DPUs as per the default and run this job.

```
datasource0 = glueContext.create_dynamic_frame.from_options(connection_type="s3",
 connection_options = {"paths": [input_path],
 "useS3ListImplementation":True,"recurse":True}, format="json")
applymapping1 = ApplyMapping.apply(frame = datasource0, mappings = [(map_spec])
datasink2 = glueContext.write_dynamic_frame.from_options(frame = applymapping1,
 connection_type = "s3", connection_options = {"path": output_path}, format =
 "parquet")
```

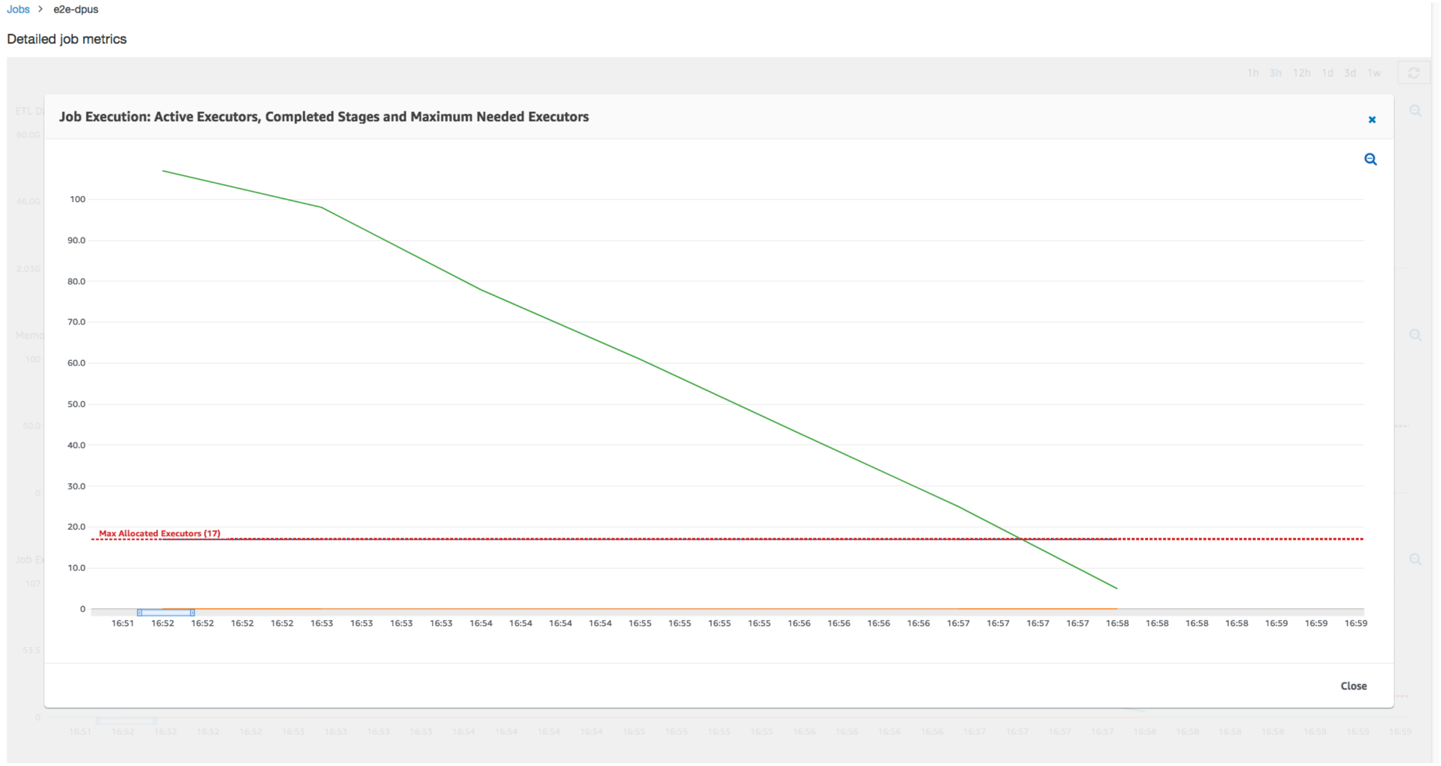
Visualize the profiled metrics on the AWS Glue console

Job run 1: In this job run we show how to find if there are under-provisioned DPUs in the cluster. The job execution functionality in AWS Glue shows the total number of actively running executors, the number of completed stages, and the number of maximum needed executors.

The number of maximum needed executors is computed by adding the total number of running tasks and pending tasks, and dividing by the tasks per executor. This result is a measure of the total number of executors required to satisfy the current load.

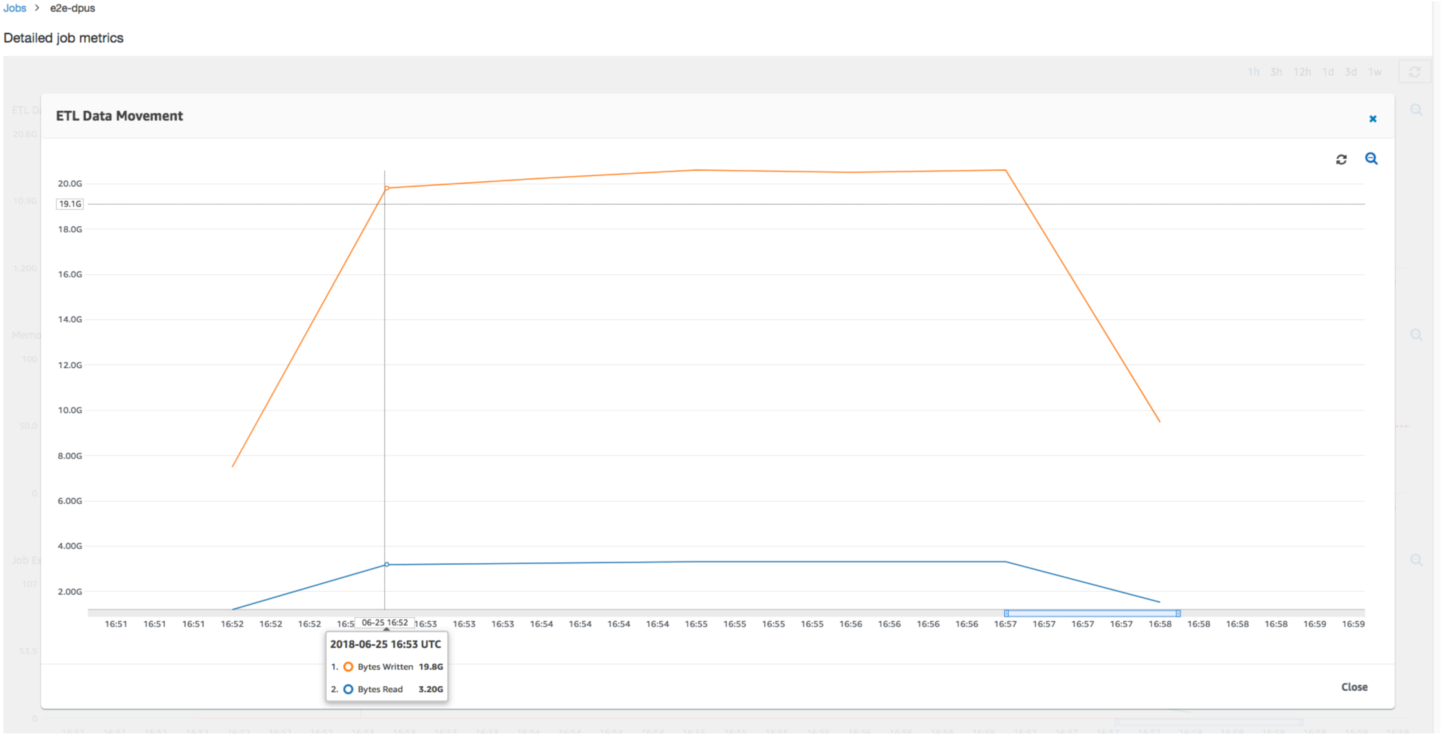
In contrast, the number of actively running executors measures how many executors are running active Apache Spark tasks. As the job progresses, the maximum needed executors can change and typically goes down towards the end of the job as the pending task queue diminishes.

The horizontal red line in the following graph shows the number of maximum allocated executors, which depends on the number of DPUs that you allocate for the job. In this case, you allocate 10 DPUs for the job run. One DPU is reserved for management. Nine DPUs run two executors each and one executor is reserved for the Spark driver. The Spark driver runs inside the primary application. So, the number of maximum allocated executors is 2*9 - 1 = 17 executors.


Jobs > e2e-dpus
Detailed job metrics
Job Execution: Active Executors, Completed Stages and Maximum Needed Executors
100
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0
Max Allocated Executors (17)
16:51
16:52
16:52
16:52
16:52
16:53
16:53
16:53
16:53
16:54
16:54
16:54
16:54
16:55
16:55
16:55
16:55
16:56
16:56
16:56
16:56
16:57
16:57
16:57
16:57
16:58
16:58
16:58
16:58
16:59
16:59
16:59
Close

As the graph shows, the number of maximum needed executors starts at 107 at the beginning of the job, whereas the number of active executors remains 17. This is the same as the number of maximum allocated executors with 10 DPUs. The ratio between the maximum needed executors and maximum allocated executors (adding 1 to both for the Spark driver) gives you the under-provisioning factor: 108/18 = 6x. You can provision 6 (under provisioning ratio) *9 (current DPU capacity - 1) + 1 DPUs = 55 DPUs to scale out the job to run it with maximum parallelism and finish faster.

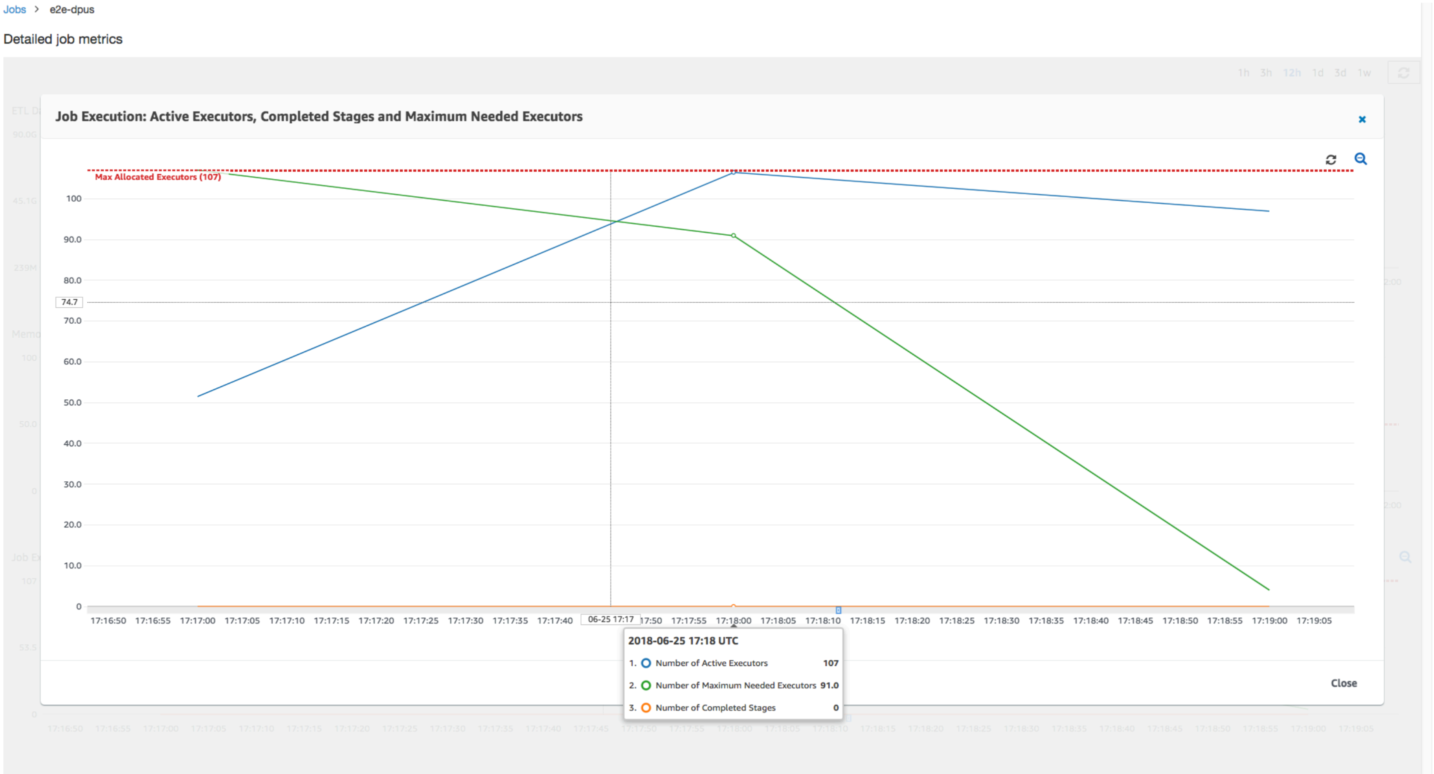
The AWS Glue console displays the detailed job metrics as a static line representing the original number of maximum allocated executors. The console computes the maximum allocated executors from the job definition for the metrics. By constrast, for detailed job run metrics, the console computes the maximum allocated executors from the job run configuration, specifically the DPUs allocated for the job run. To view metrics for an individual job run, select the job run and choose **View run metrics**.



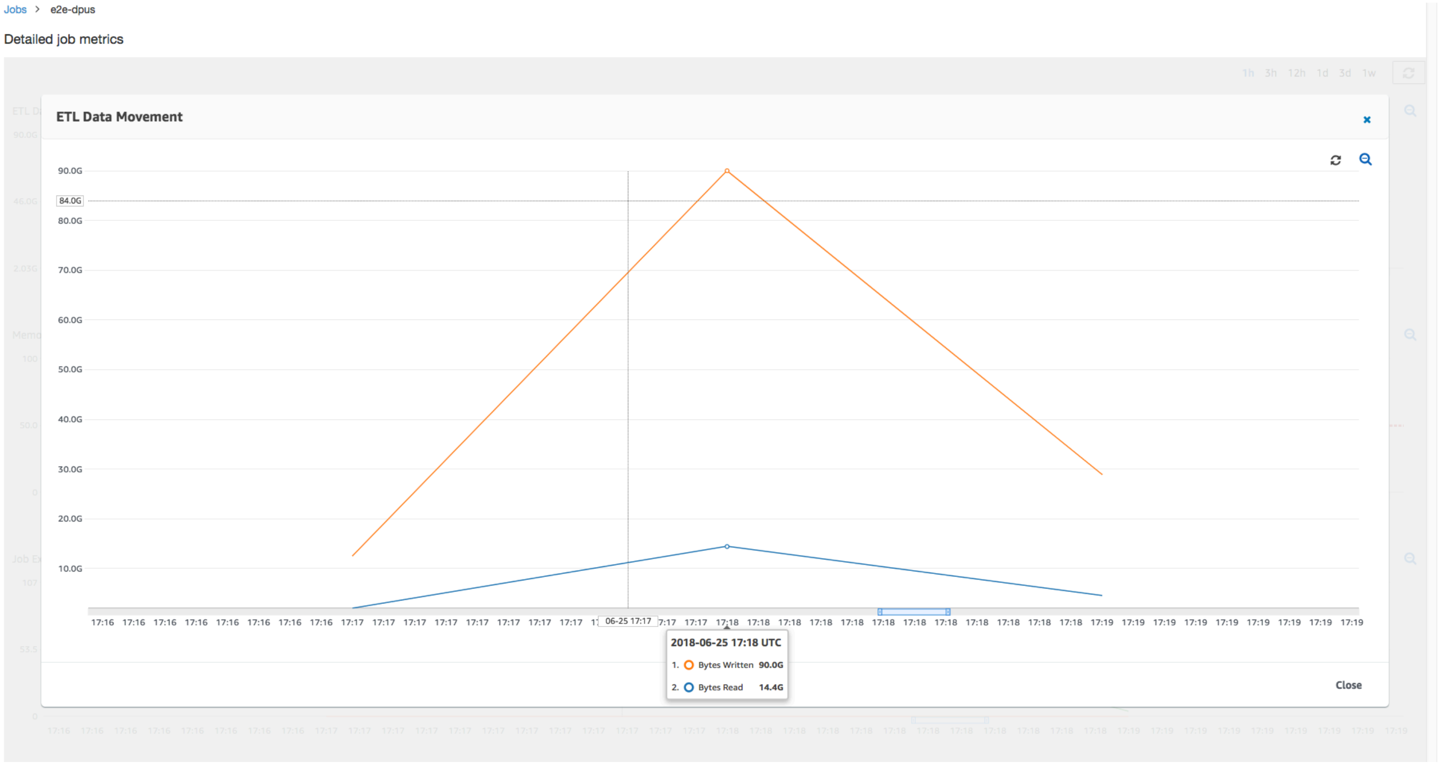
Looking at the Amazon S3 bytes read and written, notice that the job spends all six minutes streaming in data from Amazon S3 and writing it out in parallel. All the cores on the allocated DPUs are reading and writing to Amazon S3. The maximum number of needed executors being 107 also matches the number of files in the input Amazon S3 path—428. Each executor can launch four Spark tasks to process four input files (JSON gzipped).

Determine the optimal DPU capacity

Based on the results of the previous job run, you can increase the total number of allocated DPUs to 55, and see how the job performs. The job finishes in less than three minutes—half the time it required previously. The job scale-out is not linear in this case because it is a short running job. Jobs with long-lived tasks or a large number of tasks (a large number of max needed executors) benefit from a close-to-linear DPU scale-out performance speedup.


Jobs > e2e-dpus
Detailed job metrics
Job Execution: Active Executors, Completed Stages and Maximum Needed Executors
Max Allocated Executors (107)
100
90.0
80.0
74.7
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0
17:16:50
17:16:55
17:17:00
17:17:05
17:17:10
17:17:15
17:17:20
17:17:25
17:17:30
17:17:35
17:17:40
06-25 17:17
17:17:50
17:17:55
17:18:00
17:18:05
17:18:10
17:18:15
17:18:20
17:18:25
17:18:30
17:18:35
17:18:40
17:18:45
17:18:50
17:18:55
17:19:00
17:19:05
2018-06-25 17:18 UTC
1. Number of Active Executors 107
2. Number of Maximum Needed Executors 91.0
3. Number of Completed Stages 0
Close

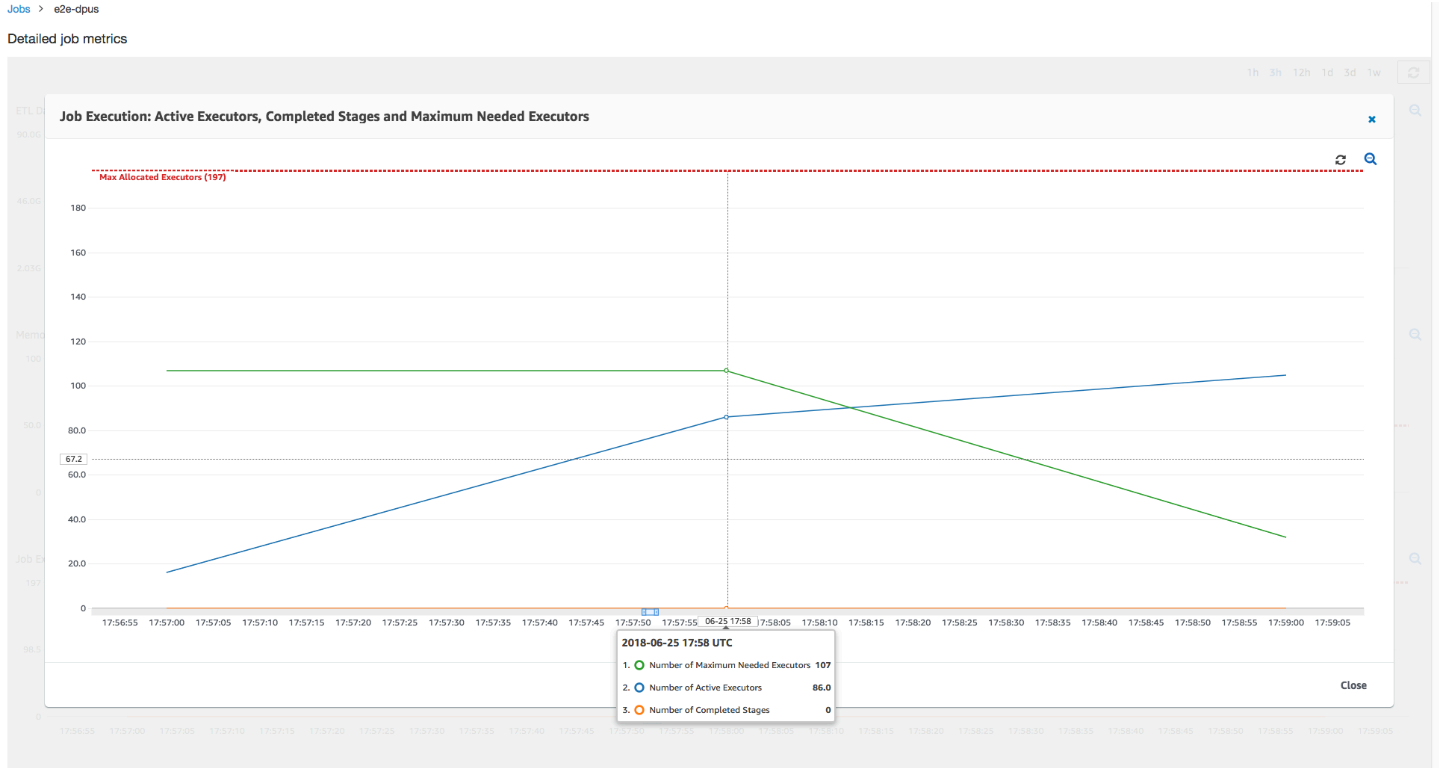
As the above image shows, the total number of active executors reaches the maximum allocated —107 executors. Similarly, the maximum needed executors is never above the maximum allocated executors. The maximum needed executors number is computed from the actively running and pending task counts, so it might be smaller than the number of active executors. This is because there can be executors that are partially or completely idle for a short period of time and are not yet decommissioned.


Jobs > e2e-dpus
Detailed job metrics
ETL Data Movement
90.0G
84.0G
80.0G
70.0G
60.0G
50.0G
40.0G
30.0G
20.0G
10.0G
17:16
17:17
17:18
17:19
06-25 17:17
2018-06-25 17:18 UTC
1. Bytes Written 90.0G
2. Bytes Read 14.4G
Close

This job run uses 6x more executors to read and write from Amazon S3 in parallel. As a result, this job run uses more Amazon S3 bandwidth for both reads and writes, and finishes faster.

Identify overprovisioned DPUs

Next, you can determine whether scaling out the job with 100 DPUs (99 * 2 = 198 executors) helps to scale out any further. As the following graph shows, the job still takes three minutes to finish. Similarly, the job does not scale out beyond 107 executors (55 DPUs configuration), and the remaining 91 executors are overprovisioned and not used at all. This shows that increasing the number of DPUs might not always improve performance, as evident from the maximum needed executors.


Jobs > e2e-dpus
Detailed job metrics
Job Execution: Active Executors, Completed Stages and Maximum Needed Executors
Max Allocated Executors (197)
180
160
140
120
100
80.0
67.2
60.0
40.0
20.0
0
17:56:55
17:57:00
17:57:05
17:57:10
17:57:15
17:57:20
17:57:25
17:57:30
17:57:35
17:57:40
17:57:45
17:57:50
17:57:55
06-25 17:58
17:58:05
17:58:10
17:58:15
17:58:20
17:58:25
17:58:30
17:58:35
17:58:40
17:58:45
17:58:50
17:58:55
17:59:00
17:59:05
2018-06-25 17:58 UTC
1. Number of Maximum Needed Executors 107
2. Number of Active Executors 86.0
3. Number of Completed Stages 0
Close

Compare time differences

The three job runs shown in the following table summarize the job execution times for 10 DPUs, 55 DPUs, and 100 DPUs. You can find the DPU capacity to improve the job execution time using the estimates you established by monitoring the first job run.

Job ID	Number of DPUs	Execution time
jr_c894524c8ef5048a4d9...	10	6 min.
jr_1a466cf2575e7ffe6856...	55	3 min.
jr_34fa1ed4c6aa9ff0a814...	100	3 min.

Generative AI troubleshooting for Apache Spark in AWS Glue

The generative AI troubleshooting for Apache Spark preview is available for jobs running on AWS Glue 4.0 and AWS Glue 5.0, and in the following AWS Regions:

US East (N. Virginia), US East (Ohio), US West (Oregon), US West (N. California), South America (São Paulo), Canada (Central), Europe (Ireland), Europe (London), Europe (Paris), Europe (Stockholm), Europe (Milan), Europe (Frankfurt), Middle East (Bahrain), Middle East (UAE), Africa (Cape Town), Asia Pacific (Tokyo), Asia Pacific (Hong Kong), Asia Pacific (Mumbai), Asia Pacific (Singapore), Asia Pacific (Jakarta), Asia Pacific (Seoul), Asia Pacific (Osaka), and Asia Pacific (Sydney).

Preview features are subject to change.

Generative AI Troubleshooting for Apache Spark jobs in AWS Glue is a new capability that helps data engineers and scientists diagnose and fix issues in their Spark applications with ease. Utilizing machine learning and generative AI technologies, this feature analyzes issues in Spark jobs and provides detailed root cause analysis along with actionable recommendations to resolve those issues.

How does Generative AI Troubleshooting for Apache Spark work?

For your failed Spark jobs, Generative AI Troubleshooting analyzes the job metadata and the precise metrics and logs associated with the error signature of your job to generate a root cause analysis, and recommends specific solutions and best practices to help address job failures.

Setting up Generative AI Troubleshooting for Apache Spark for your jobs

Configuring IAM permissions

Granting permissions to the APIs used by Spark Troubleshooting for your jobs in AWS Glue requires appropriate IAM permissions. You can obtain permissions by attaching the following custom AWS policy to your IAM identity (such as a user, role, or group).

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:StartCompletion",
        "glue:GetCompletion"
        ],
        "Resource": [
          "arn:aws:glue:*:*:completion/*",
          "arn:aws:glue:*:*:job/*"
        ]
      }
    ]
  }
```

Note

During preview, Spark Troubleshooting does not have APIs available through the AWS SDK that you can use programmatically. The following two APIs are used in the IAM policy for enabling this experience through the AWS Glue Studio Console: `StartCompletion` and `GetCompletion`.

Assigning permissions

To provide access, add permissions to your users, groups, or roles:

- For users and groups in IAM Identity Center: Create a permission set. Follow the instructions in Create a permission set in the IAM Identity Center User Guide.
- For users managed in IAM through an identity provider: Create a role for identity federation. Follow the instructions in Creating a role for a third-party identity provider (federation) in the IAM User Guide.
- For IAM users: Create a role that your user can assume. Follow the instructions in Creating a role for an IAM user in the IAM User Guide.

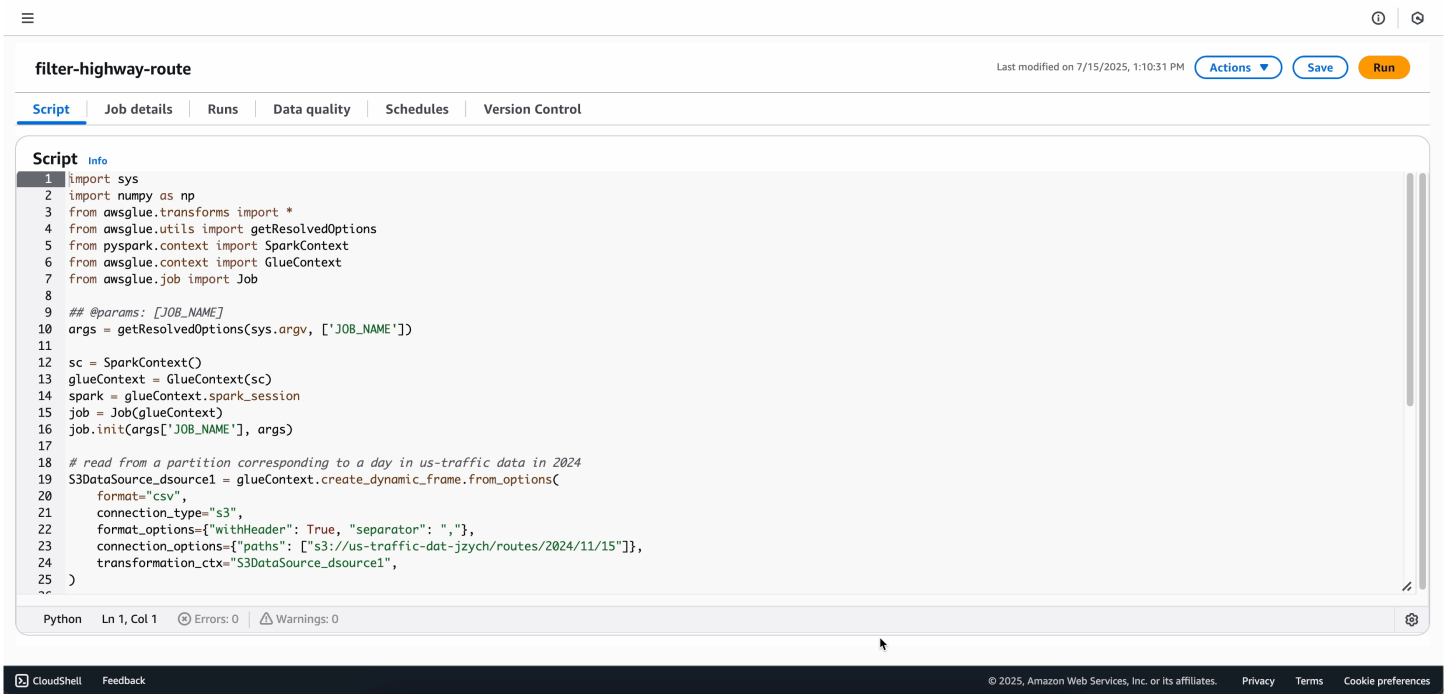
Running troubleshooting analysis from a failed job run

You can access the troubleshooting feature through multiple paths in the AWS Glue console. Here's how to get started:

Option 1: From the Jobs List page

1. Open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, choose **ETL Jobs**.

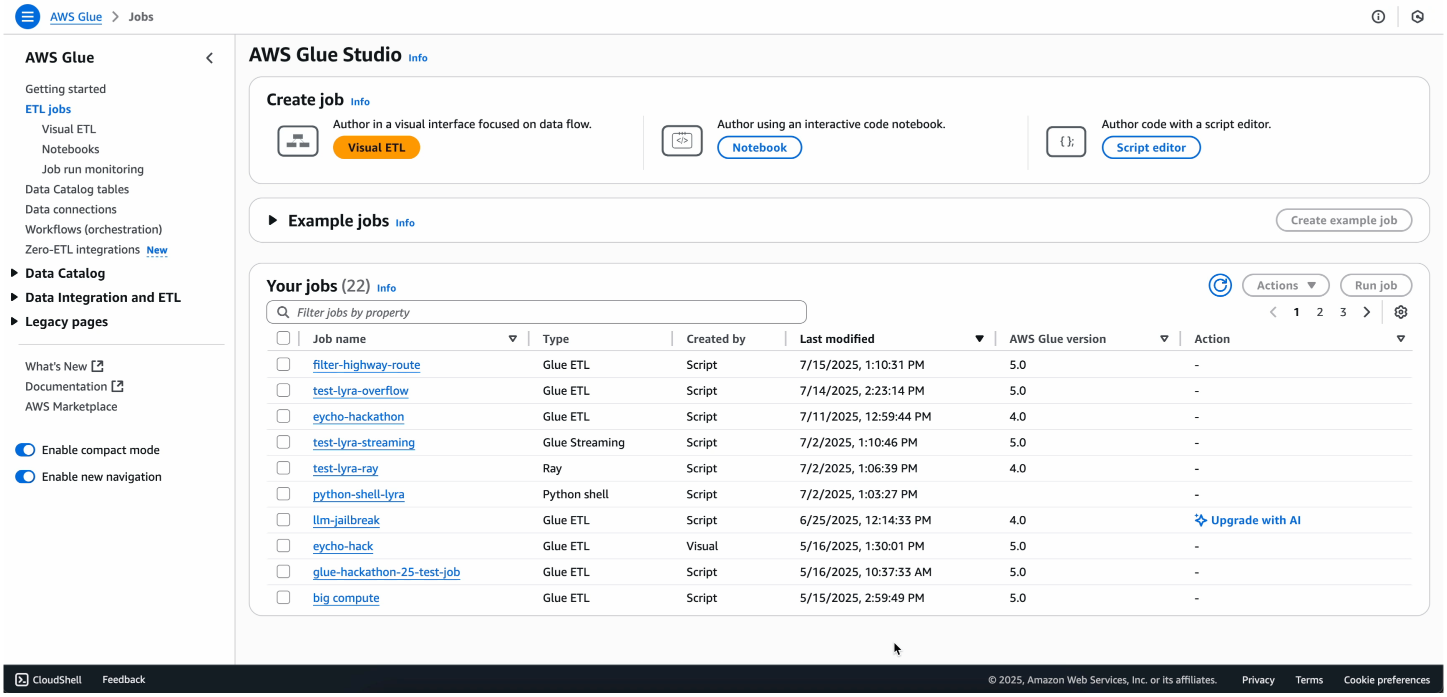
3. Locate your failed job in the jobs list.
4. Select the **Runs** tab in the job details section.
5. Click on the failed job run you want to analyze.
6. Choose **Troubleshoot with AI** to start the analysis.
7. When the troubleshooting analysis is complete, you can view the root-cause analysis and recommendations in the **Troubleshooting analysis** tab at the bottom of the screen.


filter-highway-route
Last modified on 7/15/2025, 1:10:31 PM
Actions
Save
Run
Script
Job details
Runs
Data quality
Schedules
Version Control
Script
Info
Python
Ln 1, Col 1
Errors: 0
Warnings: 0
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

```python
import sys
import numpy as np
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job

## @params: [JOB_NAME]
args = getResolvedOptions(sys.argv, ['JOB_NAME'])

sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
job.init(args['JOB_NAME'], args)

# read from a partition corresponding to a day in us-traffic data in 2024
S3DataSource_dsource1 = glueContext.create_dynamic_frame.from_options(
    format="csv",
    connection_type="s3",
    format_options={"withHeader": True, "separator": ","},
    connection_options={"paths": ["s3://us-traffic-dat-jzych/routes/2024/11/15"]},
    transformation_ctx="S3DataSource_dsource1",
)
```

Option 2: Using the Job Run Monitoring page

1. Navigate to the **Job run monitoring** page.
2. Locate your failed job run.
3. Choose the **Actions** drop-down menu.
4. Choose **Troubleshoot with AI**.


AWS Glue > Jobs
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Zero-ETL integrations New
Data Catalog
Data Integration and ETL
Legacy pages
What's New
Documentation
AWS Marketplace
Enable compact mode
Enable new navigation
AWS Glue Studio Info
Create job Info
Author in a visual interface focused on data flow.
Visual ETL
Author using an interactive code notebook.
Notebook
Author code with a script editor.
Script editor
Example jobs Info
Create example job
Your jobs (22) Info
Actions
Run job
Filter jobs by property
1 2 3
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

Job name	Type	Created by	Last modified	AWS Glue version	Action
filter-highway-route	Glue ETL	Script	7/15/2025, 1:10:31 PM	5.0	-
test-lyra-overflow	Glue ETL	Script	7/14/2025, 2:23:14 PM	5.0	-
eycho-hackathon	Glue ETL	Script	7/11/2025, 12:59:44 PM	4.0	-
test-lyra-streaming	Glue Streaming	Script	7/2/2025, 1:10:46 PM	5.0	-
test-lyra-ray	Ray	Script	7/2/2025, 1:06:39 PM	4.0	-
python-shell-lyra	Python shell	Script	7/2/2025, 1:03:27 PM		-
llm-jailbreak	Glue ETL	Script	6/25/2025, 12:14:33 PM	4.0	Upgrade with AI
eycho-hack	Glue ETL	Visual	5/16/2025, 1:30:01 PM	5.0	-
glue-hackathon-25-test-job	Glue ETL	Script	5/16/2025, 10:37:33 AM	5.0	-
big compute	Glue ETL	Script	5/15/2025, 2:59:49 PM	5.0	-

Option 3: From the Job Run Details page

1. Navigate to your failed job run's details page by either clicking **View details** on a failed run from the **Runs** tab or selecting the job run from the **Job run monitoring** page.
2. In the job run details page, find the **Troubleshooting analysis** tab.

Supported troubleshooting categories (preview)

This service focuses on three primary categories of issues that data engineers and developers frequently encounter in their Spark applications:

- **Resource setup and access errors:** When running Spark applications in AWS Glue, resource setup and access errors are among the most common yet challenging issues to diagnose. These errors often occur when your Spark application attempts to interact with AWS resources but encounters permission issues, missing resources, or configuration problems.
- **Spark driver and executor memory issues:** Memory-related errors in Apache Spark jobs can be complex to diagnose and resolve. These errors often manifest when your data processing requirements exceed the available memory resources, either on the driver node or executor nodes.
- **Spark disk capacity issues:** Storage-related errors in AWS Glue Spark jobs often emerge during shuffle operations, data spilling, or when dealing with large-scale data transformations. These

errors can be particularly tricky because they might not manifest until your job has been running for a while, potentially wasting valuable compute time and resources.

> **Note**
>
> Before implementing any suggested changes in your production environment, review the suggested changes thoroughly. The service provides recommendations based on patterns and best practices, but your specific use case might require additional considerations.

AWS Glue worker types

Overview

AWS Glue provides multiple worker types to accommodate different workload requirements, from small streaming jobs to large-scale, memory-intensive data processing tasks. This section provides comprehensive information about all available worker types, their specifications, and usage recommendations.

Worker type categories

AWS Glue offers two main categories of worker types:

- **G Worker Types**: General-purpose compute workers optimized for standard ETL workloads
- **R Worker Types**: Memory-optimized workers designed for memory-intensive Spark applications

Data Processing Units (DPUs)

The resources available on AWS Glue workers are measured in DPUs. A DPU is a relative measure of processing power that consists of 4 vCPUs of compute capacity and 16 GB of memory.

Memory-Optimized DPUs (M-DPUs): R type workers use M-DPUs, which provide double the memory allocation for a given size compared to standard DPUs. This means that while a standard DPU provides 16 GB of memory, an M-DPU in R type workers provides 32GB of memory optimized for memory-intensive Spark applications.

Available worker types

G.1X - Standard Worker

- **DPU**: 1 DPU (4 vCPUs, 16 GB memory)
- **Storage**: 94GB disk (approximately 44GB free)
- **Use Case**: Data transforms, joins, and queries - scalable and cost-effective for most jobs

G.2X - Standard Worker

- **DPU**: 2 DPU (8 vCPUs, 32 GB memory)
- **Storage**: 138GB disk (approximately 78GB free)
- **Use Case**: Data transforms, joins, and queries - scalable and cost-effective for most jobs

G.4X - Large Worker

- **DPU**: 4 DPU (16 vCPUs, 64 GB memory)
- **Storage**: 256GB disk (approximately 230GB free)
- **Use Case**: Demanding transforms, aggregations, joins, and queries

G.8X - Extra Large Worker

- **DPU**: 8 DPU (32 vCPUs, 128 GB memory)
- **Storage**: 512GB disk (approximately 485GB free)
- **Use Case**: Most demanding transforms, aggregations, joins, and queries

G.12X - Very Large Worker*

- **DPU**: 12 DPU (48 vCPUs, 192 GB memory)
- **Storage**: 768GB disk (approximately 741GB free)
- **Use Case**: Very large and resource-intensive workloads requiring significant compute capacity

G.16X - Maximum Worker*

- **DPU**: 16 DPU (64 vCPUs, 256 GB memory)

- **Storage**: 1024GB disk (approximately 996GB free)
- **Use Case**: Largest and most resource-intensive workloads requiring maximum compute capacity

R.1X - Memory-Optimized Small*

- **DPU**: 1 M-DPU (4 vCPUs, 32 GB memory)
- **Use Case**: Memory-intensive workloads with frequent out-of-memory errors or high memory-to-CPU ratio requirements

R.2X - Memory-Optimized Medium*

- **DPU**: 2 M-DPU (8 vCPUs, 64 GB memory)
- **Use Case**: Memory-intensive workloads with frequent out-of-memory errors or high memory-to-CPU ratio requirements

R.4X - Memory-Optimized Large*

- **DPU**: 4 M-DPU (16 vCPUs, 128 GB memory)
- **Use Case**: Large memory-intensive workloads with frequent out-of-memory errors or high memory-to-CPU ratio requirements

R.8X - Memory-Optimized Extra Large*

- **DPU**: 8 M-DPU (32 vCPUs, 256 GB memory)
- **Use Case**: Very large memory-intensive workloads with frequent out-of-memory errors or high memory-to-CPU ratio requirements

* You may encounter higher startup latency with these workers. To resolve the issue, try the following:

- Wait a few minutes and then submit your job again.
- Submit a new job with a reduced number of workers.
- Submit a new job using a different worker type or size.

Worker type specifications table

Worker Type Specifications

Worker Type	DPU per Node	vCPU	Memory (GB)	Disk (GB)	Free Disk Space (GB)	Spark Executors per Node
G.1X	1	4	16	94	44	1
G.2X	2	8	32	138	78	1
G.4X	4	16	64	256	230	1
G.8X	8	32	128	512	485	1
G.12X	**12**	**48**	**192**	**768**	**741**	**1**
G.16X	**16**	**64**	**256**	**1024**	**996**	**1**

Note: R worker types have memory-optimized configurations with specifications optimized for memory-intensive workloads.

Important considerations

Startup latency

⚠ **Important**

G.12X and G.16X worker types, as well as all R worker types (R.1X through R.8X), may encounter higher startup latency. To resolve the issue, try the following:

- Wait a few minutes and then submit your job again.
- Submit a new job with a reduced number of workers.
- Submit a new job using a different worker type and size.

Choosing the right worker type

For standard ETL workloads

- **G.1X or G.2X**: Most cost-effective for typical data transforms, joins, and queries
- **G.4X or G.8X**: For more demanding workloads with larger datasets

For large-scale workloads

- **G.12X**: Very large datasets requiring significant compute resources
- **G.16X**: Maximum compute capacity for the most demanding workloads

For memory-intensive workloads

- **R.1X or R.2X**: Small to medium memory-intensive jobs
- **R.4X or R.8X**: Large memory-intensive workloads with frequent OOM errors

Cost Optimization Considerations

- **Standard G workers**: Provide a balance of compute, memory and networking resources, and can be used for a variety of diverse workloads at lower cost
- **R workers**: Specialized for memory-intensive tasks with fast performance for workloads that process large data sets in memory

Best practices

Worker selection guidelines

1. **Start with standard workers** (G.1X, G.2X) for most workloads
2. **Use R workers** when experiencing frequent out-of-memory errors or workloads with memory-intensive operations like caching, shuffling, and aggregating
3. **Consider G.12X/G.16X** for compute-intensive workloads requiring maximum resources
4. **Account for capacity constraints** when using new worker types in time-sensitive workflows

Performance optimization

- Monitor CloudWatch metrics to understand resource utilization
- Use appropriate worker counts based on data size and complexity
- Consider data partitioning strategies to optimize worker efficiency

Streaming ETL jobs in AWS Glue

You can create streaming extract, transform, and load (ETL) jobs that run continuously, consume data from streaming sources like Amazon Kinesis Data Streams, Apache Kafka, and Amazon Managed Streaming for Apache Kafka (Amazon MSK). The jobs cleanse and transform the data, and then load the results into Amazon S3 data lakes or JDBC data stores.

Additionally, you can produce data for Amazon Kinesis Data Streams streams. This feature is only available when writing AWS Glue scripts. For more information, see the section called “Kinesis connections”.

By default, AWS Glue processes and writes out data in 100-second windows. This allows data to be processed efficiently and permits aggregations to be performed on data arriving later than expected. You can modify this window size to increase timeliness or aggregation accuracy. AWS Glue streaming jobs use checkpoints rather than job bookmarks to track the data that has been read.

> **Note**
>
> AWS Glue bills hourly for streaming ETL jobs while they are running.

This video discusses streaming ETL cost challenges, and cost-saving features in AWS Glue.

Creating a streaming ETL job involves the following steps:

1. For an Apache Kafka streaming source, create an AWS Glue connection to the Kafka source or the Amazon MSK cluster.
2. Manually create a Data Catalog table for the streaming source.
3. Create an ETL job for the streaming data source. Define streaming-specific job properties, and supply your own script or optionally modify the generated script.

For more information, see Streaming ETL in AWS Glue.

When creating a streaming ETL job for Amazon Kinesis Data Streams, you don't have to create an AWS Glue connection. However, if there is a connection attached to the AWS Glue streaming ETL job that has Kinesis Data Streams as a source, then a virtual private cloud (VPC) endpoint to Kinesis is required. For more information, see Creating an interface endpoint in the *Amazon VPC User Guide*. When specifying a Amazon Kinesis Data Streams stream in another account, you must setup the roles and policies to allow cross-account access. For more information, see Example: Read From a Kinesis Stream in a Different Account.

AWS Glue streaming ETL jobs can auto-detect compressed data, transparently decompress the streaming data, perform the usual transformations on the input source, and load to the output store.

AWS Glue supports auto-decompression for the following compression types given the input format:

Compression type	Avro file	Avro datum	JSON	CSV	Grok
BZIP2	Yes	Yes	Yes	Yes	Yes
GZIP	No	Yes	Yes	Yes	Yes
SNAPPY	Yes (raw Snappy)	Yes (framed Snappy)	Yes (framed Snappy)	Yes (framed Snappy)	Yes (framed Snappy)
XZ	Yes	Yes	Yes	Yes	Yes
ZSTD	Yes	No	No	No	No
DEFLATE	Yes	Yes	Yes	Yes	Yes

Topics

- Creating an AWS Glue connection for an Apache Kafka data stream
- Creating a Data Catalog table for a streaming source
- Notes and restrictions for Avro streaming sources
- Applying grok patterns to streaming sources

- Defining job properties for a streaming ETL job
- Streaming ETL notes and restrictions

Creating an AWS Glue connection for an Apache Kafka data stream

To read from an Apache Kafka stream, you must create an AWS Glue connection.

To create an AWS Glue connection for a Kafka source (Console)

1. Open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, under **Data catalog**, choose **Connections**.
3. Choose **Add connection**, and on the **Set up your connection's properties** page, enter a connection name.

   > **Note**
   >
   > For more information about specifying connection properties, see AWS Glue connection properties..

4. For **Connection type**, choose **Kafka**.
5. For **Kafka bootstrap servers URLs**, enter the host and port number for the bootstrap brokers for your Amazon MSK cluster or Apache Kafka cluster. Use only Transport Layer Security (TLS) endpoints for establishing the initial connection to the Kafka cluster. Plaintext endpoints are not supported.

   The following is an example list of hostname and port number pairs for an Amazon MSK cluster.

   ```
   myserver1.kafka.us-east-1.amazonaws.com:9094,myserver2.kafka.us-
   east-1.amazonaws.com:9094,
   myserver3.kafka.us-east-1.amazonaws.com:9094
   ```

   For more information about getting the bootstrap broker information, see Getting the Bootstrap Brokers for an Amazon MSK Cluster in the *Amazon Managed Streaming for Apache Kafka Developer Guide*.

6. If you want a secure connection to the Kafka data source, select **Require SSL connection**, and for **Kafka private CA certificate location**, enter a valid Amazon S3 path to a custom SSL certificate.

   For an SSL connection to self-managed Kafka, the custom certificate is mandatory. It's optional for Amazon MSK.

   For more information about specifying a custom certificate for Kafka, see the section called "SSL connection properties".
7. Use AWS Glue Studio or the AWS CLI to specify a Kafka client authentication method. To access AWS Glue Studio select **AWS Glue** from the **ETL** menu in the left navigation pane.

   For more information about Kafka client authentication methods, see AWS Glue Kafka connection properties for client authentication .
8. Optionally enter a description, and then choose **Next**.
9. For an Amazon MSK cluster, specify its virtual private cloud (VPC), subnet, and security group. The VPC information is optional for self-managed Kafka.
10. Choose **Next** to review all connection properties, and then choose **Finish**.

For more information about AWS Glue connections, see Connecting to data.

AWS Glue Kafka connection properties for client authentication

SASL/GSSAPI (Kerberos) authentication

Choosing this authentication method will allow you to specify Kerberos properties.

Kerberos Keytab

Choose the location of the keytab file. A keytab stores long-term keys for one or more principals. For more information, see MIT Kerberos Documentation: Keytab .

Kerberos krb5.conf file

Choose the krb5.conf file. This contains the default realm (a logical network, similar to a domain, that defines a group of systems under the same KDC) and the location of the KDC server. For more information, see MIT Kerberos Documentation: krb5.conf .

Kerberos principal and Kerberos service name

Enter the Kerberos principal and service name. For more information, see MIT Kerberos Documentation: Kerberos principal .

SASL/SCRAM-SHA-512 authentication

Choosing this authentication method will allow you to specify authentication credentials.

AWS Secrets Manager

Search for your token in the Search box by typing the name or ARN.

Provider username and password directly

Search for your token in the Search box by typing the name or ARN.

SSL client authentication

Choosing this authentication method allows you to select the location of the Kafka client keystore by browsing Amazon S3. Optionally, you can enter the Kafka client keystore password and Kafka client key password.

IAM authentication

This authentication method does not require any additional specifications and is only applicable when the Streaming source is MSK Kafka.

SASL/PLAIN authentication

Choosing this authentication method allows you to specify authentication credentials.

Creating a Data Catalog table for a streaming source

A Data Catalog table that specifies source data stream properties, including the data schema can be manually created for a streaming source. This table is used as the data source for the streaming ETL job.

If you don't know the schema of the data in the source data stream, you can create the table without a schema. Then when you create the streaming ETL job, you can turn on the AWS Glue schema detection function. AWS Glue determines the schema from the streaming data.

Use the AWS Glue console, the AWS Command Line Interface (AWS CLI), or the AWS Glue API to create the table. For information about creating a table manually with the AWS Glue console, see the section called “Creating tables”.

> **Note**
>
> You can't use the AWS Lake Formation console to create the table; you must use the AWS Glue console.

Also consider the following information for streaming sources in Avro format or for log data that you can apply Grok patterns to.

- the section called "Notes and restrictions for Avro streaming sources"
- the section called "Applying grok patterns to streaming sources"

Topics

- Kinesis data source
- Kafka data source
- AWS Glue Schema Registry table source

Kinesis data source

When creating the table, set the following streaming ETL properties (console).

Type of Source

Kinesis

For a Kinesis source in the same account:

Region

The AWS Region where the Amazon Kinesis Data Streams service resides. The Region and Kinesis stream name are together translated to a Stream ARN.

Example: https://kinesis.us-east-1.amazonaws.com

Kinesis stream name

Stream name as described in Creating a Stream in the *Amazon Kinesis Data Streams Developer Guide*.

For a Kinesis source in another account, refer to this example to set up the roles and policies to allow cross-account access. Configure these settings:

Stream ARN

The ARN of the Kinesis data stream that the consumer is registered with. For more information, see Amazon Resource Names (ARNs) and AWS Service Namespaces in the *AWS General Reference*.

Assumed Role ARN

The Amazon Resource Name (ARN) of the role to assume.

Session name (optional)

An identifier for the assumed role session.

Use the role session name to uniquely identify a session when the same role is assumed by different principals or for different reasons. In cross-account scenarios, the role session name is visible to, and can be logged by the account that owns the role. The role session name is also used in the ARN of the assumed role principal. This means that subsequent cross-account API requests that use the temporary security credentials will expose the role session name to the external account in their AWS CloudTrail logs.

To set streaming ETL properties for Amazon Kinesis Data Streams (AWS Glue API or AWS CLI)

- To set up streaming ETL properties for a Kinesis source in the same account, specify the `streamName` and `endpointUrl` parameters in the `StorageDescriptor` structure of the `CreateTable` API operation or the `create_table` CLI command.

```
"StorageDescriptor": {
 "Parameters": {
  "typeOfData": "kinesis",
  "streamName": "sample-stream",
  "endpointUrl": "https://kinesis.us-east-1.amazonaws.com"
 }
 ...
}
```

  Or, specify the `streamARN`.

Example

```
"StorageDescriptor": {
 "Parameters": {
  "typeOfData": "kinesis",
  "streamARN": "arn:aws:kinesis:us-east-1:123456789:stream/sample-stream"
 }
 ...
}
```

- To set up streaming ETL properties for a Kinesis source in another account, specify the `streamARN`, `awsSTSRoleARN` and `awsSTSSessionName` (optional) parameters in the `StorageDescriptor` structure in the `CreateTable` API operation or the `create_table` CLI command.

```
"StorageDescriptor": {
 "Parameters": {
  "typeOfData": "kinesis",
  "streamARN": "arn:aws:kinesis:us-east-1:123456789:stream/sample-stream",
  "awsSTSRoleARN": "arn:aws:iam::123456789:role/sample-assume-role-arn",
  "awsSTSSessionName": "optional-session"
 }
 ...
}
```

Kafka data source

When creating the table, set the following streaming ETL properties (console).

Type of Source

Kafka

For a Kafka source:

Topic name

Topic name as specified in Kafka.

Connection

An AWS Glue connection that references a Kafka source, as described in the section called “Creating a connection for a Kafka data stream”.

AWS Glue Schema Registry table source

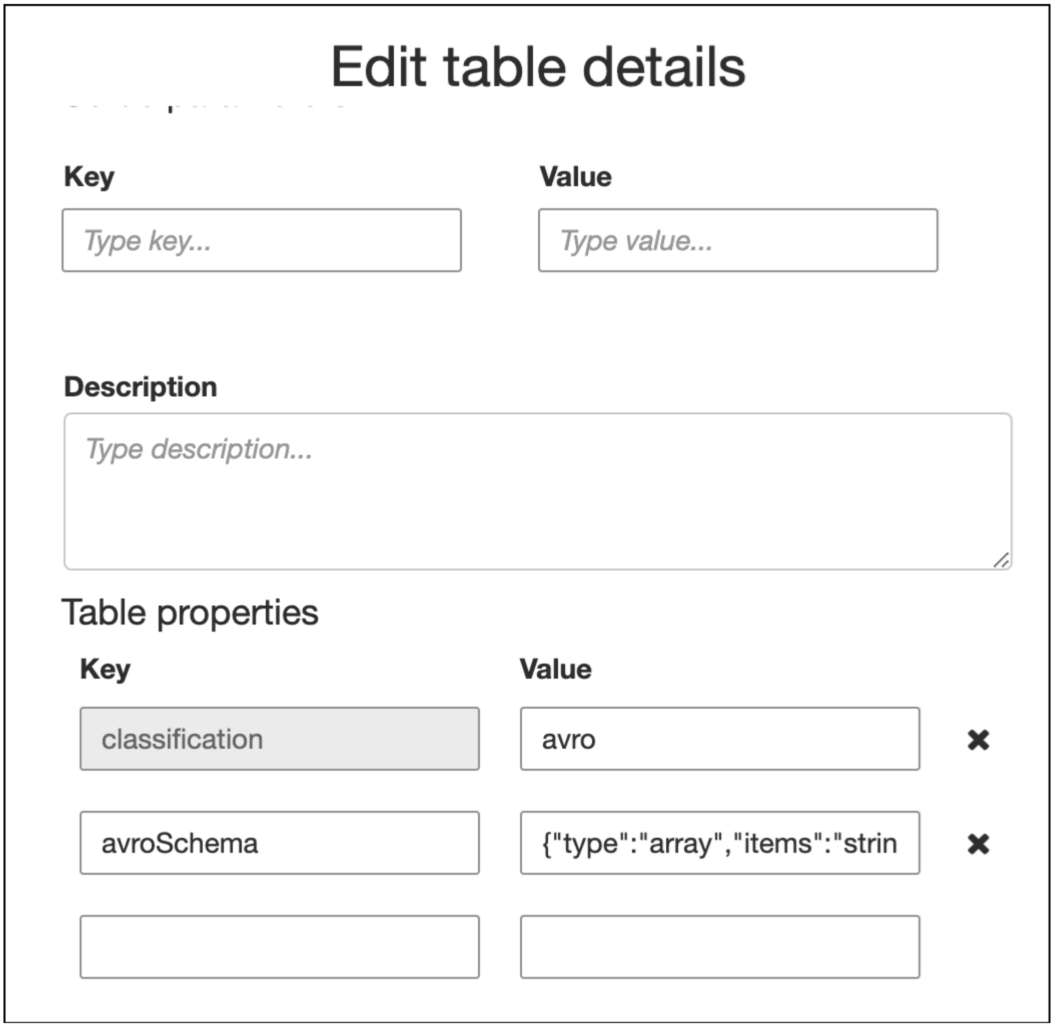
To use AWS Glue Schema Registry for streaming jobs, follow the instructions at Use case: AWS Glue Data Catalog to create or update a Schema Registry table.

Currently, AWS Glue Streaming supports only Glue Schema Registry Avro format with schema inference set to `false`.

Notes and restrictions for Avro streaming sources

The following notes and restrictions apply for streaming sources in the Avro format:

- When schema detection is turned on, the Avro schema must be included in the payload. When turned off, the payload should contain only data.
- Some Avro data types are not supported in dynamic frames. You can't specify these data types when defining the schema with the **Define a schema** page in the create table wizard in the AWS Glue console. During schema detection, unsupported types in the Avro schema are converted to supported types as follows:
  - `EnumType => StringType`
  - `FixedType => BinaryType`
  - `UnionType => StructType`
- If you define the table schema using the **Define a schema** page in the console, the implied root element type for the schema is `record`. If you want a root element type other than `record`, for example `array` or `map`, you can't specify the schema using the **Define a schema** page. Instead you must skip that page and specify the schema either as a table property or within the ETL script.
  - To specify the schema in the table properties, complete the create table wizard, edit the table details, and add a new key-value pair under **Table properties**. Use the key `avroSchema`, and enter a schema JSON object for the value, as shown in the following screenshot.



Edit table details

Key: Type key...

Value: Type value...

Description: Type description...

Table properties

Key	Value
classification	avro
avroSchema	{"type":"array","items":"strin

- To specify the schema in the ETL script, modify the `datasource0` assignment statement and add the `avroSchema` key to the `additional_options` argument, as shown in the following Python and Scala examples.

  Python

  ```
  SCHEMA_STRING = ‘{"type":"array","items":"string"}’
  datasource0 = glueContext.create_data_frame.from_catalog(database =
   "database", table_name = "table_name", transformation_ctx = "datasource0",
   additional_options = {"startingPosition": "TRIM_HORIZON", "inferSchema":
   "false", "avroSchema": SCHEMA_STRING})
  ```

  Scala

  ```
  val SCHEMA_STRING = """{"type":"array","items":"string"}"""
  val datasource0 = glueContext.getCatalogSource(database = "database", tableName
   = "table_name", redshiftTmpDir = "", transformationContext = "datasource0",
  ```

```
additionalOptions = JsonOptions(s"""{"startingPosition": "TRIM_HORIZON", 
"inferSchema": "false", "avroSchema":"$SCHEMA_STRING"}""")).getDataFrame()
```

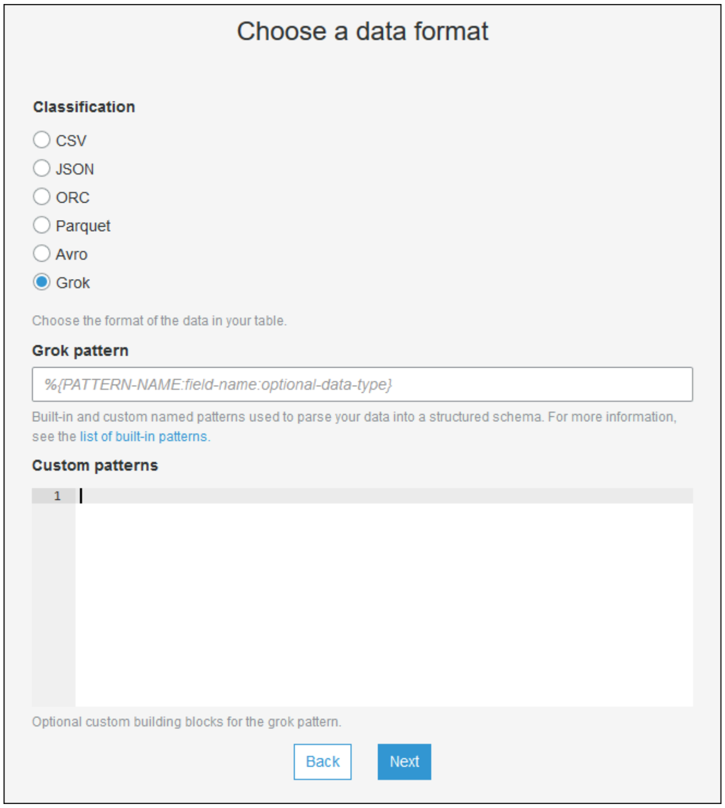
Applying grok patterns to streaming sources

You can create a streaming ETL job for a log data source and use Grok patterns to convert the logs to structured data. The ETL job then processes the data as a structured data source. You specify the Grok patterns to apply when you create the Data Catalog table for the streaming source.

For information about Grok patterns and custom pattern string values, see Writing grok custom classifiers.

To add grok patterns to the Data Catalog table (console)

- Use the create table wizard, and create the table with the parameters specified in the section called “Creating a Data Catalog table for a streaming source”. Specify the data format as Grok, fill in the **Grok pattern** field, and optionally add custom patterns under **Custom patterns (optional)**.

  Choose a data format

  Classification

  ○ CSV
  ○ JSON
  ○ ORC
  ○ Parquet
  ○ Avro
  ● Grok

  Choose the format of the data in your table.

  Grok pattern

  %{PATTERN-NAME:field-name:optional-data-type}

  Built-in and custom named patterns used to parse your data into a structured schema. For more information, see the list of built-in patterns.

  Custom patterns

  Optional custom building blocks for the grok pattern.

  Press **Enter** after each custom pattern.

To add grok patterns to the Data Catalog table (AWS Glue API or AWS CLI)

- Add the `GrokPattern` parameter and optionally the `CustomPatterns` parameter to the `CreateTable` API operation or the `create_table` CLI command.

  ```
   "Parameters": {
  ...
      "grokPattern": "string",
      "grokCustomPatterns": "string",
  ...
  },
  ```

  Express `grokCustomPatterns` as a string and use "\n" as the separator between patterns.

  The following is an example of specifying these parameters.

  Example

  ```
  "parameters": {
  ...
      "grokPattern": "%{USERNAME:username} %{DIGIT:digit:int}",
      "grokCustomPatterns": "digit \d",
  ...
  }
  ```

Defining job properties for a streaming ETL job

When you define a streaming ETL job in the AWS Glue console, provide the following streams-specific properties. For descriptions of additional job properties, see Defining job properties for Spark jobs.

IAM role

Specify the AWS Identity and Access Management (IAM) role that is used for authorization to resources that are used to run the job, access streaming sources, and access target data stores.

For access to Amazon Kinesis Data Streams, attach the `AmazonKinesisFullAccess` AWS managed policy to the role, or attach a similar IAM policy that permits more fine-grained access. For sample policies, see Controlling Access to Amazon Kinesis Data Streams Resources Using IAM.

For more information about permissions for running jobs in AWS Glue, see Identity and access management for AWS Glue.

Type

Choose **Spark streaming**.

AWS Glue version

The AWS Glue version determines the versions of Apache Spark, and Python or Scala, that are available to the job. Choose a selection that specifies the version of Python or Scala available to the job. AWS Glue Version 2.0 with Python 3 support is the default for streaming ETL jobs.

Maintenance window

Specifies a window where a streaming job can be restarted. See the section called “Maintenance windows”.

Job timeout

Optionally enter a duration in minutes. The default value is blank.

- Streaming jobs must have a timeout value less than 7 days or 10080 minutes.
- When the value is left blank, the job will be restarted after 7 days, if you have not set up a maintenance window. If you have set up a maintenance window, the job will be restarted during the maintenance window after 7 days.

Data source

Specify the table that you created in the section called “Creating a Data Catalog table for a streaming source”.

Data target

Do one of the following:

- Choose **Create tables in your data target** and specify the following data target properties.

  Data store

  Choose Amazon S3 or JDBC.

  Format

  Choose any format. All are supported for streaming.
- Choose **Use tables in the data catalog and update your data target**, and choose a table for a JDBC data store.

Output schema definition

Do one of the following:

- Choose **Automatically detect schema of each record** to turn on schema detection. AWS Glue determines the schema from the streaming data.
- Choose **Specify output schema for all records** to use the Apply Mapping transform to define the output schema.

Script

Optionally supply your own script or modify the generated script to perform operations that the Apache Spark Structured Streaming engine supports. For information on the available operations, see Operations on streaming DataFrames/Datasets.

Streaming ETL notes and restrictions

Keep in mind the following notes and restrictions:

- Auto-decompression for AWS Glue streaming ETL jobs is only available for the supported compression types. Also note the following:
  - Framed Snappy refers to the official framing format for Snappy.
  - Deflate is supported in Glue version 3.0, not Glue version 2.0.
- When using schema detection, you cannot perform joins of streaming data.
- AWS Glue streaming ETL jobs do not support the Union data type for AWS Glue Schema Registry with Avro format.
- Your ETL script can use AWS Glue's built-in transforms and the transforms native to Apache Spark Structured Streaming. For more information, see Operations on streaming DataFrames/Datasets on the Apache Spark website or AWS Glue PySpark transforms reference.
- AWS Glue streaming ETL jobs use checkpoints to keep track of the data that has been read. Therefore, a stopped and restarted job picks up where it left off in the stream. If you want to reprocess data, you can delete the checkpoint folder referenced in the script.
- Job bookmarks aren't supported.
- To use enhanced fan-out feature of Kinesis Data Streams in your job, consult the section called "Using enhanced fan-out in Kinesis streaming jobs".
- If you use a Data Catalog table created from AWS Glue Schema Registry, when a new schema version becomes available, to reflect the new schema, you need to do the following:

1. Stop the jobs associated with the table.
2. Update the schema for the Data Catalog table.
3. Restart the jobs associated with the table.

Record matching with AWS Lake Formation FindMatches

> **Note**
>
> Record matching is currently unavailable in the following Regions in the AWS Glue console: Middle East (UAE), Europe (Spain), Asia Pacific (Jakarta), and Europe (Zurich).

AWS Lake Formation provides machine learning capabilities to create custom transforms to cleanse your data. There is currently one available transform named FindMatches. The FindMatches transform enables you to identify duplicate or matching records in your dataset, even when the records do not have a common unique identifier and no fields match exactly. This will not require writing any code or knowing how machine learning works. FindMatches can be useful in many different problems, such as:

- **Matching customers**: Linking customer records across different customer databases, even when many customer fields do not match exactly across the databases (e.g. different name spelling, address differences, missing or inaccurate data, etc).
- **Matching products**: Matching products in your catalog against other product sources, such as product catalog against a competitor's catalog, where entries are structured differently.
- **Improving fraud detection**: Identifying duplicate customer accounts, determining when a newly created account is (or might be) a match for a previously known fraudulent user.
- **Other matching problems**: Match addresses, movies, parts lists, etc etc. In general, if a human being could look at your database rows and determine that they were a match, there is a really good chance that the FindMatches transform can help you.

You can create these transforms when you create a job. The transform that you create is based on a source data store schema and example data from the source data set that you label (we call this process "teaching" a transform). The records that you label must be present in the source dataset. In this process we generate a file which you label and then upload back which the transform would in a manner learn from. After you teach your transform, you can call it from your Spark-based AWS Glue job (PySpark or Scala Spark) and use it in other scripts with a compatible source data store.

After the transform is created, it is stored in AWS Glue. On the AWS Glue console, you can manage the transforms that you create. In the navigation pane under **Data Integration and ETL**, **Data classification tools > Record Matching**, you can edit and continue to teach your machine learning transform. For more information about managing transforms on the console, see Working with machine learning transforms.

> **Note**
>
> AWS Glue version 2.0 FindMatches jobs use the Amazon S3 bucket `aws-glue-temp-<accountID>-<region>` to store temporary files while the transform is processing data. You can delete this data after the run has completed, either manually or by setting an Amazon S3 Lifecycle rule.

Types of machine learning transforms

You can create machine learning transforms to cleanse your data. You can call these transforms from your ETL script. Your data passes from transform to transform in a data structure called a *DynamicFrame*, which is an extension to an Apache Spark SQL `DataFrame`. The `DynamicFrame` contains your data, and you reference its schema to process your data.

The following types of machine learning transforms are available:

Find matches

Finds duplicate records in the source data. You teach this machine learning transform by labeling example datasets to indicate which rows match. The machine learning transform learns which rows should be matches the more you teach it with example labeled data. Depending on how you configure the transform, the output is one of the following:

- A copy of the input table plus a `match_id` column filled in with values that indicate matching sets of records. The `match_id` column is an arbitrary identifier. Any records which have the same `match_id` have been identified as matching to each other. Records with different `match_id`'s do not match.
- A copy of the input table with duplicate rows removed. If multiple duplicates are found, then the record with the lowest primary key is kept.

Find incremental matches

> The Find matches transform can also be configured to find matches across the existing and incremental frames and return as output a column containing a unique ID per match group.
>
> For more information, see: Finding incremental matches

Using the FindMatches transform

You can use the `FindMatches` transform to find duplicate records in the source data. A labeling file is generated or provided to help teach the transform.

> **Note**
>
> Currently, `FindMatches` transforms that use a custom encryption key aren't supported in the following Regions:
>
> - Asia Pacific (Osaka) - `ap-northeast-3`

To get started with the FindMatches transform, you can follow the steps below. For a more advanced and detailed example, see the **AWS Big Data Blog**: Harmonize data using AWS Glue and AWS Lake Formation FindMatches ML to build a customer 360 view .

Getting started using the Find Matches transform

Follow these steps to get started with the `FindMatches` transform:

1. Create a table in the AWS Glue Data Catalog for the source data that is to be cleaned. For information about how to create a crawler, see Working with Crawlers on the AWS Glue Console.

   If your source data is a text-based file such as a comma-separated values (CSV) file, consider the following:

   - Keep your input record CSV file and labeling file in separate folders. Otherwise, the AWS Glue crawler might consider them as multiple parts of the same table and create tables in the Data Catalog incorrectly.
   - Unless your CSV file includes ASCII characters only, ensure that UTF-8 without BOM (byte order mark) encoding is used for the CSV files. Microsoft Excel often adds a BOM in the beginning of UTF-8 CSV files. To remove it, open the CSV file in a text editor, and resave the file as **UTF-8 without BOM**.

2. On the AWS Glue console, create a job, and choose the **Find matches** transform type.

> ⚠ **Important**
>
> The data source table that you choose for the job can't have more than 100 columns.

3. Tell AWS Glue to generate a labeling file by choosing **Generate labeling file**. AWS Glue takes the first pass at grouping similar records for each `labeling_set_id` so that you can review those groupings. You label matches in the `label` column.
   - If you already have a labeling file, that is, an example of records that indicate matching rows, upload the file to Amazon Simple Storage Service (Amazon S3). For information about the format of the labeling file, see Labeling file format. Proceed to step 4.
4. Download the labeling file and label the file as described in the Labeling section.
5. Upload the corrected labelled file. AWS Glue runs tasks to teach the transform how to find matches.

   On the **Machine learning transforms** list page, choose the **History** tab. This page indicates when AWS Glue performs the following tasks:
   - **Import labels**
   - **Export labels**
   - **Generate labels**
   - **Estimate quality**
6. To create a better transform, you can iteratively download, label, and upload the labelled file. In the initial runs, a lot more records might be mismatched. But AWS Glue learns as you continue to teach it by verifying the labeling file.
7. Evaluate and tune your transform by evaluating performance and results of finding matches. For more information, see Tuning machine learning transforms in AWS Glue.

Labeling

When `FindMatches` generates a labeling file, records are selected from your source table. Based on previous training, `FindMatches` identifies the most valuable records to learn from.

The act of *labeling* is editing a labeling file (we suggest using a spreadsheet such as Microsoft Excel) and adding identifiers, or labels, into the `label` column that identifies matching and nonmatching records. It is important to have a clear and consistent definition of a match in your source data.

`FindMatches` learns from which records you designate as matches (or not) and uses your decisions to learn how to find duplicate records.

When a labeling file is generated by `FindMatches`, approximately 100 records are generated. These 100 records are typically divided into 10 *labeling sets*, where each labeling set is identified by a unique `labeling_set_id` generated by `FindMatches`. Each labeling set should be viewed as a separate labeling task independent of the other labeling sets. Your task is to identify matching and non-matching records within each labeling set.

Tips for editing labeling files in a spreadsheet

When editing the labeling file in a spreadsheet application, consider the following:

- The file might not open with column fields fully expanded. You might need to expand the `labeling_set_id` and `label` columns to see content in those cells.
- If the primary key column is a number, such as a `long` data type, the spreadsheet might interpret it as a number and change the value. This key value must be treated as text. To correct this problem, format all the cells in the primary key column as **Text data**.

Labeling file format

The labeling file that is generated by AWS Glue to teach your `FindMatches` transform uses the following format. If you generate your own file for AWS Glue, it must follow this format as well:

- It is a comma-separated values (CSV) file.
- It must be encoded in `UTF-8`. If you edit the file using Microsoft Windows, it might be encoded with `cp1252`.
- It must be in an Amazon S3 location to pass it to AWS Glue.
- Use a moderate number of rows for each labeling task. 10–20 rows per task are recommended, although 2–30 rows per task are acceptable. Tasks larger than 50 rows are not recommended and may cause poor results or system failure.
- If you have already-labeled data consisting of pairs of records labeled as a "match" or a "no-match", this is fine. These labeled pairs can be represented as labeling sets of size 2. In this case label both records with, for instance, a letter "A" if they match, but label one as "A" and one as "B" if they do not match.

Note

Because it has additional columns, the labeling file has a different schema from a file that contains your source data. Place the labeling file in a different folder from any transform input CSV file so that the AWS Glue crawler does not consider it when it creates tables in the Data Catalog. Otherwise, the tables created by the AWS Glue crawler might not correctly represent your data.

- The first two columns (`labeling_set_id`, `label`) are required by AWS Glue. The remaining columns must match the schema of the data that is to be processed.
- For each `labeling_set_id`, you identify all matching records by using the same label. A label is a unique string placed in the `label` column. We recommend using labels that contain simple characters, such as A, B, C, and so on. Labels are case sensitive and are entered in the `label` column.
- Rows that contain the same `labeling_set_id` and the same label are understood to be labeled as a match.
- Rows that contain the same `labeling_set_id` and a different label are understood to be labeled as *not* a match
- Rows that contain a different `labeling_set_id` are understood to be conveying no information for or against matching.

The following is an example of labeling the data:

labeling_set_id	label	first_name	last_name	Birthday
ABC123	A	John	Doe	04/01/1980
ABC123	B	Jane	Smith	04/03/1980
ABC123	A	Johnny	Doe	04/01/1980
ABC123	A	Jon	Doe	04/01/1980
DEF345	A	Richard	Jones	12/11/1992
DEF345	A	Rich	Jones	11/12/1992

labeling_set_id	label	first_name	last_name	Birthday
DEF345	B	Sarah	Jones	12/11/1992
DEF345	C	Richie	Jones Jr.	05/06/2017
DEF345	B	Sarah	Jones-Walker	12/11/1992
GHI678	A	Robert	Miller	1/3/1999
GHI678	A	Bob	Miller	1/3/1999
XYZABC	A	William	Robinson	2/5/2001
XYZABC	B	Andrew	Robinson	2/5/1971

- In the above example we identify John/Johnny/Jon Doe as being a match and we teach the system that these records do not match Jane Smith. Separately, we teach the system that Richard and Rich Jones are the same person, but that these records are not a match to Sarah Jones/Jones-Walker and Richie Jones Jr.
- As you can see, the scope of the labels is limited to the `labeling_set_id`. So labels do not cross `labeling_set_id` boundaries. For example, a label "A" in `labeling_set_id` 1 does not have any relation to label "A" in `labeling_set_id` 2.
- If a record does not have any matches within a labeling set, then assign it a unique label. For instance, Jane Smith does not match any record in labeling set ABC123, so it is the only record in that labeling set with the label of B.
- The labeling set "GHI678" shows that a labeling set can consist of just two records which are given the same label to show that they match. Similarly, "XYZABC" shows two records given different labels to show that they do not match.
- Note that sometimes a labeling set may contain no matches (that is, you give every record in the labeling set a different label) or a labeling set might all be "the same" (you gave them all the same label). This is okay as long as your labeling sets collectively contain examples of records that are and are not "the same" by your criteria.

Important

Confirm that the IAM role that you pass to AWS Glue has access to the Amazon S3 bucket that contains the labeling file. By convention, AWS Glue policies grant permission to

Amazon S3 buckets or folders whose names are prefixed with **aws-glue-**. If your labeling files are in a different location, add permission to that location in the IAM role.

Tuning machine learning transforms in AWS Glue

You can tune your machine learning transforms in AWS Glue to improve the results of your data-cleansing jobs to meet your objectives. To improve your transform, you can teach it by generating a labeling set, adding labels, and then repeating these steps several times until you get your desired results. You can also tune by changing some machine learning parameters.

For more information about machine learning transforms, see Record matching with AWS Lake Formation FindMatches.

Topics

- Machine learning measurements
- Deciding between precision and recall
- Deciding Between accuracy and cost
- Estimating the quality of matches using match confidence scores
- Teaching the Find Matches transform

Machine learning measurements

To understand the measurements that are used to tune your machine learning transform, you should be familiar with the following terminology:

True positive (TP)

A match in the data that the transform correctly found, sometimes called a *hit*.

True negative (TN)

A nonmatch in the data that the transform correctly rejected.

False positive (FP)

A nonmatch in the data that the transform incorrectly classified as a match, sometimes called a *false alarm*.

False negative (FN)

A match in the data that the transform didn't find, sometimes called a *miss*.

For more information about the terminology that is used in machine learning, see Confusion matrix in Wikipedia.

To tune your machine learning transforms, you can change the value of the following measurements in the **Advanced properties** of the transform.

- **Precision** measures how well the transform finds true positives among the total number of records that it identifies as positive (true positives and false positives). For more information, see Precision and recall in Wikipedia.
- **Recall** measures how well the transform finds true positives from the total records in the source data. For more information, see Precision and recall in Wikipedia.
- **Accuracy** measures how well the transform finds true positives and true negatives. Increasing accuracy requires more machine resources and cost. But it also results in increased recall. For more information, see Accuracy and precision in Wikipedia.
- **Cost** measures how many compute resources (and thus money) are consumed to run the transform.

Deciding between precision and recall

Each `FindMatches` transform contains a `precision-recall` parameter. You use this parameter to specify one of the following:

- If you are more concerned about the transform falsely reporting that two records match when they actually don't match, then you should emphasize *precision*.
- If you are more concerned about the transform failing to detect records that really do match, then you should emphasize *recall*.

You can make this trade-off on the AWS Glue console or by using the AWS Glue machine learning API operations.

When to favor precision

Favor precision if you are more concerned about the risk that `FindMatches` results in a pair of records matching when they don't actually match. To favor precision, choose a *higher* precision-recall trade-off value. With a higher value, the `FindMatches` transform requires more evidence to decide that a pair of records should be matched. The transform is tuned to bias toward saying that records do not match.

For example, suppose that you're using `FindMatches` to detect duplicate items in a video catalog, and you provide a higher precision-recall value to the transform. If your transform incorrectly detects that *Star Wars: A New Hope* is the same as *Star Wars: The Empire Strikes Back*, a customer who wants *A New Hope* might be shown *The Empire Strikes Back*. This would be a poor customer experience.

However, if the transform fails to detect that *Star Wars: A New Hope* and *Star Wars: Episode IV —A New Hope* are the same item, the customer might be confused at first but might eventually recognize them as the same. It would be a mistake, but not as bad as the previous scenario.

When to favor recall

Favor recall if you are more concerned about the risk that the `FindMatches` transform results might fail to detect a pair of records that actually do match. To favor recall, choose a *lower* precision-recall trade-off value. With a lower value, the `FindMatches` transform requires less evidence to decide that a pair of records should be matched. The transform is tuned to bias toward saying that records do match.

For example, this might be a priority for a security organization. Suppose that you are matching customers against a list of known defrauders, and it is important to determine whether a customer is a defrauder. You are using `FindMatches` to match the defrauder list against the customer list. Every time `FindMatches` detects a match between the two lists, a human auditor is assigned to verify that the person is, in fact, a defrauder. Your organization might prefer to choose recall over precision. In other words, you would rather have the auditors manually review and reject some cases when the customer is not a defrauder than fail to identify that a customer is, in fact, on the defrauder list.

How to favor both precision and recall

The best way to improve both precision and recall is to label more data. As you label more data, the overall accuracy of the `FindMatches` transform improves, thus improving both precision and recall. Nevertheless, even with the most accurate transform, there is always a gray area where you need to experiment with favoring precision or recall, or choose a value in the middle.

Deciding Between accuracy and cost

Each `FindMatches` transform contains an `accuracy-cost` parameter. You can use this parameter to specify one of the following:

- If you are more concerned with the transform accurately reporting that two records match, then you should emphasize *accuracy*.
- If you are more concerned about the cost or speed of running the transform, then you should emphasize *lower cost*.

You can make this trade-off on the AWS Glue console or by using the AWS Glue machine learning API operations.

When to favor accuracy

Favor accuracy if you are more concerned about the risk that the `find matches` results won't contain matches. To favor accuracy, choose a *higher* accuracy-cost trade-off value. With a higher value, the `FindMatches` transform requires more time to do a more thorough search for correctly matching records. Note that this parameter doesn't make it less likely to falsely call a nonmatching record pair a match. The transform is tuned to bias towards spending more time finding matches.

When to favor cost

Favor cost if you are more concerned about the cost of running the `find matches` transform and less about how many matches are found. To favor cost, choose a *lower* accuracy-cost trade-off value. With a lower value, the `FindMatches` transform requires fewer resources to run. The transform is tuned to bias towards finding fewer matches. If the results are acceptable when favoring lower cost, use this setting.

How to favor both accuracy and lower cost

It takes more machine time to examine more pairs of records to determine whether they might be matches. If you want to reduce cost without reducing quality, here are some steps you can take:

- Eliminate records in your data source that you aren't concerned about matching.
- Eliminate columns from your data source that you are sure aren't useful for making a match/no-match decision. A good way of deciding this is to eliminate columns that you don't think affect your own decision about whether a set of records is “the same.”

Estimating the quality of matches using match confidence scores

Match confidence scores provide an estimate of the quality of matches found by FindMatches to distinguish between matched records in which the machine learning model is highly confident, uncertain, or unlikely. A match confidence score will be between 0 and 1, where a higher score means higher similarity. Examining match confidence scores lets you distinguish between clusters of matches in which the system is highly confident (which you may decide to merge), clusters about which the system is uncertain (which you may decide to have reviewed by a human), and clusters that the system deems to be unlikely (which you may decide to reject).

You may want to adjust your training data in situations where you see a high match confidence score, but determine there are not matches, or where you see a low score but determine there are, in fact, matches.

Confidence scores are particularly useful when there are large sized industrial datasets, where it is infeasible to review every FindMatches decision.

Match confidence scores are available in AWS Glue version 2.0 or later.

Generating match confidence scores

You can generate match confidence scores by setting the Boolean value of `computeMatchConfidenceScores` to True when calling the `FindMatches` or `FindIncrementalMatches` API.

AWS Glue adds a new `column match_confidence_score` to the output.

Match scoring examples

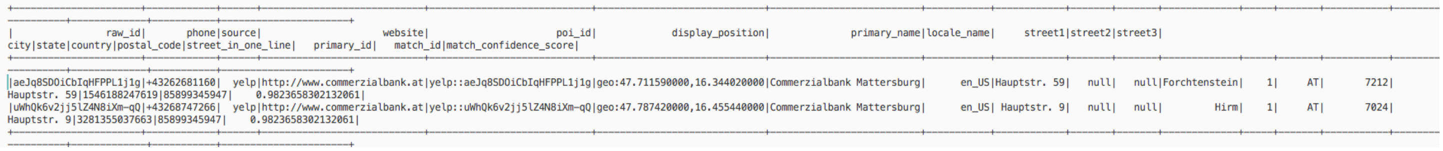
For example, consider the following matched records:

Score >= 0.9

Summary of matched records:

```
  primary_id  |   match_id  | match_confidence_score

3281355037663    85899345947   0.9823658302132061
1546188247619    85899345947   0.9823658302132061
```

Details:



```
|               raw_id|       phone|source|                    website|                   poi_id|            display_position|            primary_name|locale_name|     street1|street2|street3|
city|state|country|postal_code|street_in_one_line|   primary_id|   match_id|match_confidence_score|
|aeJq8SDOiCbIqHFPPL1j1g|+43262681160|  yelp|http://www.commerzialbank.at|yelp::aeJq8SDOiCbIqHFPPL1j1g|geo:47.711590000,16.344020000|Commerzialbank Mattersburg|      en_US|Hauptstr. 59|   null|   null|Forchtenstein|    1|     AT|       7212|
Hauptstr. 59|1546188247619|85899345947|    0.9823658302132061|
|uWhQk6v2jj5lZ4N8iXm-qQ|+43268747266|  yelp|http://www.commerzialbank.at|yelp::uWhQk6v2jj5lZ4N8iXm-qQ|geo:47.787420000,16.455440000|Commerzialbank Mattersburg|      en_US| Hauptstr. 9|   null|   null|         Hirm|    1|     AT|       7024|
Hauptstr. 9|3281355037663|85899345947|    0.9823658302132061|
```

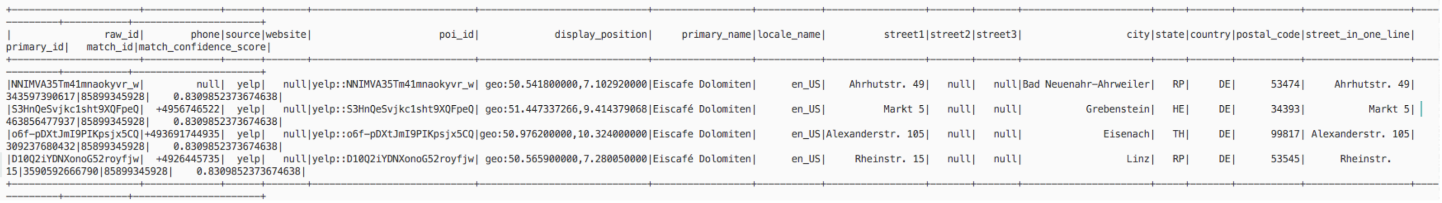
From this example, we can see that two records are very similar and share `display_position`, `primary_name`, and `street name`.

Score >= 0.8 and score < 0.9

Summary of matched records:

```
  primary_id  |   match_id  | match_confidence_score

309237680432     85899345928   0.8309852373674638
3590592666790    85899345928   0.8309852373674638
343597390617     85899345928   0.8309852373674638
249108124906     85899345928   0.8309852373674638
463856477937     85899345928   0.8309852373674638
```

Details:



```
|               raw_id|       phone|source|website|                    poi_id|            display_position|     primary_name|locale_name|         street1|street2|street3|                 city|state|country|postal_code|street_in_one_line|
primary_id|   match_id|match_confidence_score|
|NNIMVA35Tm41mnaokyvr_w|        null|  yelp|   null|yelp::NNIMVA35Tm41mnaokyvr_w| geo:50.541800000,7.102920000|Eiscafe Dolomiten|      en_US|   Ahrhutstr. 49|   null|   null|Bad Neuenahr-Ahrweiler|   RP|     DE|      53474|     Ahrhutstr. 49|
343597390617|85899345928|    0.8309852373674638|
|S3HnQeSvjkc1sht9XQFpeQ|  +4956746522|  yelp|   null|yelp::S3HnQeSvjkc1sht9XQFpeQ| geo:51.447337266,9.414379068|Eiscafé Dolomiten|      en_US|         Markt 5|   null|   null|           Grebenstein|   HE|     DE|      34393|           Markt 5|
463856477937|85899345928|    0.8309852373674638|
|o6f-pDXtJmI9PIKpsjx5CQ|+493691744935|  yelp|   null|yelp::o6f-pDXtJmI9PIKpsjx5CQ|geo:50.976200000,10.324000000|Eiscafe Dolomiten|      en_US|Alexanderstr. 105|   null|   null|              Eisenach|   TH|     DE|      99817| Alexanderstr. 105|
309237680432|85899345928|    0.8309852373674638|
|D10Q2iYDNXonoG52royfjw|  +4926445735|  yelp|   null|yelp::D10Q2iYDNXonoG52royfjw| geo:50.565900000,7.280050000|Eiscafé Dolomiten|      en_US|    Rheinstr. 15|   null|   null|                  Linz|   RP|     DE|      53545|      Rheinstr.
15|3590592666790|85899345928|    0.8309852373674638|
```

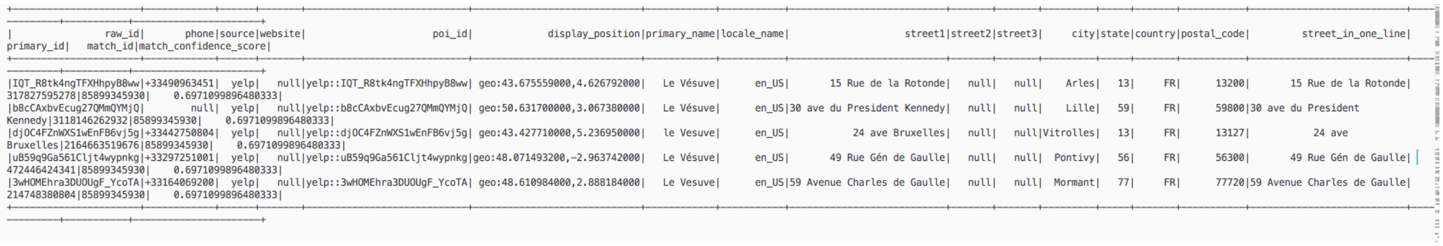
From this example, we can see that these records share the same `primary_name`, and `country`.

Score >= 0.6 and score < 0.7

Summary of matched records:

```
  primary_id  |   match_id  | match_confidence_score

2164663519676    85899345930   0.6971099896480333
 317827595278    85899345930   0.6971099896480333
 472446424341    85899345930   0.6971099896480333
3118146262932    85899345930   0.6971099896480333
 214748380804    85899345930   0.6971099896480333
```

Details:



```
+----------------------+------------+------+-------+----------------------------+----------------------------+------------+-----------+----------------------------+-------+-------+---------+-----+-------+-----------+----------------------------+-
---------+------------+----------------------+
|                raw_id|       phone|source|website|                      poi_id|            display_position|primary_name|locale_name|                     street1|street2|street3|     city|state|country|postal_code|          street_in_one_line|
primary_id|   match_id|match_confidence_score|
+----------------------+------------+------+-------+----------------------------+----------------------------+------------+-----------+----------------------------+-------+-------+---------+-----+-------+-----------+----------------------------+-
---------+------------+----------------------+
|IQT_R8tk4ngTFXHhpyB8ww|+33490963451|  yelp|   null|yelp::IQT_R8tk4ngTFXHhpyB8ww| geo:43.675559000,4.626792000|   Le Vésuve|      en_US|        15 Rue de la Rotonde|   null|   null|    Arles|   13|     FR|      13200|        15 Rue de la Rotonde|
317827595278|85899345930|    0.6971099896480333|
|b8cCAxbvEcug27QMmQYMjQ|        null|  yelp|   null|yelp::b8cCAxbvEcug27QMmQYMjQ| geo:50.631700000,3.067380000|   Le Vésuve|      en_US|30 ave du President Kennedy|   null|   null|    Lille|   59|     FR|      59800|30 ave du President
Kennedy|3118146262932|85899345930|    0.6971099896480333|
|djOC4FZnWXS1wEnFB6vj5g|+33442750804|  yelp|   null|yelp::djOC4FZnWXS1wEnFB6vj5g| geo:43.427710000,5.236950000|   le Vesuve|      en_US|            24 ave Bruxelles|   null|   null|Vitrolles|   13|     FR|      13127|            24 ave
Bruxelles|2164663519676|85899345930|    0.6971099896480333|
|uB59q9Ga561Cljt4wypnkg|+33297251001|  yelp|   null|yelp::uB59q9Ga561Cljt4wypnkg|geo:48.071493200,-2.963742000|   Le Vésuve|      en_US|        49 Rue Gén de Gaulle|   null|   null|  Pontivy|   56|     FR|      56300|        49 Rue Gén de Gaulle|
472446424341|85899345930|    0.6971099896480333|
|3wHOMEhra3DUOUgF_YcoTA|+33164069200|  yelp|   null|yelp::3wHOMEhra3DUOUgF_YcoTA| geo:48.610984000,2.888184000|   Le Vesuve|      en_US|59 Avenue Charles de Gaulle|   null|   null|  Mormant|   77|     FR|      77720|59 Avenue Charles de Gaulle|
214748380804|85899345930|    0.6971099896480333|
+----------------------+------------+------+-------+----------------------------+----------------------------+------------+-----------+----------------------------+-------+-------+---------+-----+-------+-----------+----------------------------+-
---------+------------+----------------------+
```

From this example, we can see that these records share only the same `primary_name`.

For more information, see:

- Step 5: Add and run a job with your machine learning transform
- PySpark: FindMatches class
- PySpark: FindIncrementalMatches class
- Scala: FindMatches class
- Scala: FindIncrementalMatches class

Teaching the Find Matches transform

Each `FindMatches` transform must be taught what should be considered a match and what should not be considered a match. You teach your transform by adding labels to a file and uploading your choices to AWS Glue.

You can orchestrate this labeling on the AWS Glue console or by using the AWS Glue machine learning API operations.

How many times should I add labels? How many labels do I need?

The answers to these questions are mostly up to you. You must evaluate whether `FindMatches` is delivering the level of accuracy that you need and whether you think the extra labeling effort is worth it for you. The best way to decide this is to look at the “Precision,” “Recall,” and “Area under the precision recall curve” metrics that you can generate when you choose **Estimate quality** on the AWS Glue console. After you label more sets of tasks, rerun these metrics and verify whether they have improved. If, after labeling a few sets of tasks, you don't see improvement on the metric that you are focusing on, the transform quality might have reached a plateau.

Why are both true positive and true negative labels needed?

The `FindMatches` transform needs both positive and negative examples to learn what you think is a match. If you are labeling `FindMatches`-generated training data (for example, using the **I do**

not have labels option), `FindMatches` tries to generate a set of “label set ids” for you. Within each task, you give the same “label” to some records and different “labels” to other records. In other words, the tasks generally are not either all the same or all different (but it's okay if a particular task is all “the same” or all “not the same”).

If you are teaching your `FindMatches` transform using the **Upload labels from S3** option, try to include both examples of matching and nonmatching records. It's acceptable to have only one type. These labels help you build a more accurate `FindMatches` transform, but you still need to label some records that you generate using the **Generate labeling file** option.

How can I enforce that the transform matches exactly as I taught it?

The `FindMatches` transform learns from the labels that you provide, so it might generate records pairs that don't respect the provided labels. To enforce that the `FindMatches` transform respects your labels, select **EnforceProvidedLabels** in **FindMatchesParameter**.

What techniques can you use when an ML transform identifies items as matches that are not true matches?

You can use the following techniques:

- Increase the `precisionRecallTradeoff` to a higher value. This eventually results in finding fewer matches, but it should also break up your big cluster when it reaches a high enough value.
- Take the output rows corresponding to the incorrect results and reformat them as a labeling set (removing the `match_id` column and adding a `labeling_set_id` and `label` column). If necessary, break up (subdivide) into multiple labeling sets to ensure that the labeler can keep each labeling set in mind while assigning labels. Then, correctly label the matching sets and upload the label file and append it to your existing labels. This might teach your transformer enough about what it is looking for to understand the pattern.
- (Advanced) Finally, look at that data to see if there is a pattern that you can detect that the system is not noticing. Preprocess that data using standard AWS Glue functions to *normalize* the data. Highlight what you want the algorithm to learn from by separating data that you know to be differently important into their own columns. Or construct combined columns from columns whose data you know to be related.

Working with machine learning transforms

You can use AWS Glue to create custom machine learning transforms that can be used to cleanse your data. You can use these transforms when you create a job on the AWS Glue console.

For information about how to create a machine learning transform, see Record matching with AWS Lake Formation FindMatches.

Topics

- Transform properties
- Adding and editing machine learning transforms
- Viewing transform details
- Teach transforms using labels

Transform properties

To view an existing machine learning transform, sign in to the AWS Management Console, and open the AWS Glue console at https://console.aws.amazon.com/glue/. In the navigation pane under **Data Integration and ETL**, choose **Data classification tools > Record Matching**.

The properties for each transform:

Transform name

The unique name you gave the transform when you created it.

ID

A unique identifier of the transform.

Label count

The number of labels in the labeling file that was provided to help teach the transform.

Status

Indicates whether the transform is **Ready** or **Needs training.** To run a machine learning transform successfully in a job, it must be **Ready**.

Created

The date the transform was created.

Modified

The date the transform was last updated.

Description

The description supplied for the transform, if one was provided.

AWS Glue version

The version of AWS Glue used.

Run ID

The unique name you gave the transform when you created it.

Task type

The type of machine learning transform; for example, **Find matching records**.

Status

Indicates the status of the task run. Possible statuses include:

- Starting
- Running
- Stopping
- Stopped
- Succeeded
- Failed
- Timeout

Error

If the status is Failed, an error message is displayed describing the reason for the failure.

Adding and editing machine learning transforms

You can view, delete, set up and teach, or tune a transform on the AWS Glue console. Select the check box next to the transform in the list, choose **Action**, and then choose the action that you want to take.

Creating a new ML transform

To add a new machine learning transform, choose **Create transform**. Follow the instructions in the **Add job** wizard. For more information, see Record matching with AWS Lake Formation FindMatches.

Step 1. Set transform properties.

1. Enter the name and description (optional).

2. Optionally, set security configuration. See Using data encryption with machine learning transforms.
3. Optionally, set Task execution settings. Task execution settings allow you to customize how the task is run. Select the Worker type, number of workers, task timeout (in minutes), the number of retries, and the AWS Glue version.
4. Optionally, set Tags. Tags are labels that you can assign to an AWS resource. Each tag consists of a key and an optional value. Tags can be used to search and filter your resource or track your AWS costs.

Step 2. Choose table and primary key.

1. Choose the AWS Glue Catalog database and table.
2. Choose a primary key from the selected table. The primary key column typically contains a unique identifier for every record in the data source.

Step 3. Select tuning options.

1. For **Recall vs. precision**, choose the tuning value to tune the transform to favor recall or precision. By default, **Balanced** is selected, but you can choose to favor recall or favor precision, or choose **Custom** and enter a value between 0.0 and 1.0 (inclusive).
2. For **Lower cost vs. accuracy**, choose the tuning value to favor lower cost or accuracy, or choose **Custom** and enter a value between 0.0 and 1.0 (inclusive).
3. For **Match enforcement**, choose **Force output to match labels** if you want to teach the ML transform by forcing the output to match the labels used.

Step 4. Review and create.

1. Review the options for steps 1 – 3.
2. Choose **Edit** for any step that needs to be modified. Choose **Create transform** to complete the create transform wizard.

Using data encryption with machine learning transforms

When adding a machine learning transform to AWS Glue, you can optionally specify a security configuration that is associated with the data source or data target. If the Amazon S3 bucket used to store the data is encrypted with a security configuration, specify the same security configuration when creating the transform.

You can also choose to use server-side encryption with AWS KMS (SSE-KMS) to encrypt the model and labels to prevent unauthorized persons from inspecting it. If you choose this option, you're prompted to choose the AWS KMS key by name, or you can choose **Enter a key ARN**. If you choose to enter the ARN for the KMS key, a second field appears where you can enter the KMS key ARN.

> **Note**
>
> Currently, ML transforms that use a custom encryption key aren't supported in the following Regions:
>
> - Asia Pacific (Osaka) - `ap-northeast-3`

Viewing transform details

Viewing transform properties

The **Transform properties** page includes attributes of your transform. It shows you the details about the transform definition, including the following:

- **Transform name** shows the name of the transform.
- **Type** lists the type of the transform.
- **Status** displays whether the transform is ready to be used in a script or job.
- **Force output to match labels** displays whether the transform forces the output to match the labels provided by the user.
- **Spark version** is related to the AWS Glue version you chose in the **Task run properties** when adding the transform. AWS Glue 1.0 and Spark 2.4 is recommended for most customers. For more information, see AWS Glue Versions.

History, Estimate quality and Tags tabs

Transform details include the information that you defined when you created the transform. To view the details of a transform, select the transform in the **Machine learning transforms** list, and review the information on the following tabs:

- History

- Estimate quality
- Tags

History

The **History** tab shows your transform task run history. Several types of tasks are run to teach a transform. For each task, the run metrics include the following:

- **Run ID** is an identifier created by AWS Glue for each run of this task.
- **Task type** shows the type of task run.
- **Status** shows the success of each task listed with the most recent run at the top.
- **Error** shows the details of an error message if the run was not successful.
- **Start time** shows the date and time (local time) that the task started.
- **End time** shows the date and time (local time) that the task ended.
- **Logs** links to the logs written to `stdout` for this job run.

  The **Logs** link takes you to Amazon CloudWatch Logs. There you can view the details about the tables that were created in the AWS Glue Data Catalog and any errors that were encountered. You can manage your log retention period on the CloudWatch console. The default log retention is `Never Expire`. For more information about how to change the retention period, see Change Log Data Retention in CloudWatch Logs in the *Amazon CloudWatch Logs User Guide*.
- **Label file** shows a link to Amazon S3 for a generated labeling file.

Estimate quality

The **Estimate quality** tab shows the metrics that you use to measure the quality of the transform. Estimates are calculated by comparing the transform match predictions using a subset of your labeled data against the labels you have provided. These estimates are approximate. You can invoke an **Estimate quality** task run from this tab.

The **Estimate quality** tab shows the metrics from the last **Estimate quality** run including the following properties:

- **Area under the Precision-Recall curve** is a single number estimating the upper bound of the overall quality of the transform. It is independent of the choice made for the precision-recall parameter. Higher values indicate that you have a more attractive precision-recall tradeoff.

- **Precision** estimates how often the transform is correct when it predicts a match.
- **Recall upper limit** estimates that for an actual match, how often the transform predicts the match.
- **F1** estimates the transform's accuracy between 0 and 1, where 1 is the best accuracy. For more information, see F1 score in Wikipedia.
- The **Column importance** table show the column names and importance score for each column. Column importance helps you understand how columns contribute to your model, by identifying which columns in your records are being used the most to do the matching. This data may prompt you to add to or change your labelset to raise or lower the importance of columns.

  The Importance column provides a numerical score for each column, as a decimal not greater than 1.0.

For information about understanding quality estimates versus true quality, see Quality estimates versus end-to-end (true) quality.

For more information about tuning your transform, see Tuning machine learning transforms in AWS Glue.

Quality estimates versus end-to-end (true) quality

AWS Glue estimates the quality of your transform by presenting the internal machine-learned model with a number of pairs of records that you provided matching labels for but that the model has not seen before. These quality estimates are a function of the quality of the machine-learned model (which is influenced by the number of records that you label to “teach” the transform). The end-to-end, or *true* recall (which is not automatically calculated by the `ML transform`) is also influenced by the `ML transform` filtering mechanism that proposes a wide variety of possible matches to the machine-learned model.

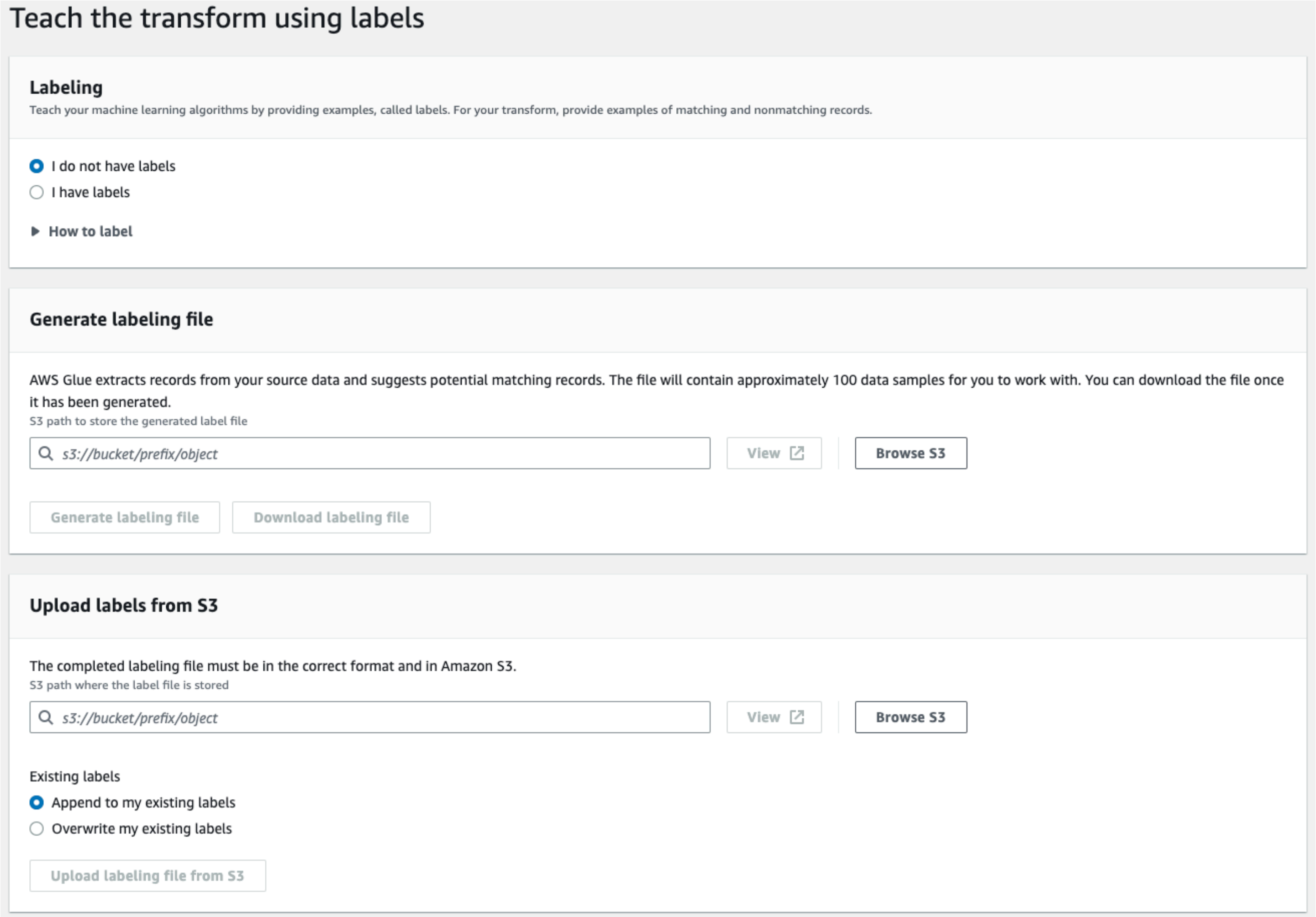
You can tune this filtering method primarily by specifying the **Lower Cost-Accuracy** tuning value. As the tuning value gets closer to favor **Accuracy**, the system does a more thorough and expensive search for pairs of records that might be matches. More pairs of records are fed to your machine-learned model, and your `ML transform`'s end-to-end or true recall gets closer to the estimated recall metric. As a result, changes in the end-to-end quality of your matches as a result of changes in the cost/accuracy tradeoff for your matches will typically not be reflected in the quality estimate.

Tags

Tags are labels that you can assign to an AWS resource. Each tag consists of a key and an optional value. Tags can be used to search and filter your resource or track your AWS costs.

Teach transforms using labels

You can teach your ML transform using labels (examples) by choosing **Teach transform** from the ML transform details page. When you teach your machine learning algorithm by providing examples (called labels), you can choose existing labels to use, or create a labeling file.



Teach the transform using labels

Labeling

Teach your machine learning algorithms by providing examples, called labels. For your transform, provide examples of matching and nonmatching records.

● I do not have labels
○ I have labels

▶ How to label

Generate labeling file

AWS Glue extracts records from your source data and suggests potential matching records. The file will contain approximately 100 data samples for you to work with. You can download the file once it has been generated.

S3 path to store the generated label file: s3://bucket/prefix/object

Upload labels from S3

The completed labeling file must be in the correct format and in Amazon S3.

S3 path where the label file is stored: s3://bucket/prefix/object

Existing labels
● Append to my existing labels
○ Overwrite my existing labels

- **Labeling** – If you have labels, choose **I have labels**. If you do not have labels, you can still continue with the next step in generating a labeling file.
- **Generate labeling file** – AWS Glue extracts records from your source data and suggest potential matching records. You choose the Amazon S3 bucket to store the generated label file. Choose **Generate labeling file** to start the process. When done, choose **Download labeling file**. The downloaded file will have a column for labels where you can fill in the labels.

- **Upload labels from Amazon S3** – Choose the completed labeling file from the Amazon S3 bucket where the label file is stored. Then, choose to either append the labels to your existing labels or to overwrite your existing labels. Choose **Upload labeling file from Amazon S3**.

Tutorial: Creating a machine learning transform with AWS Glue

This tutorial guides you through the actions to create and manage a machine learning (ML) transform using AWS Glue. Before using this tutorial, you should be familiar with using the AWS Glue console to add crawlers and jobs and edit scripts. You should also be familiar with finding and downloading files on the Amazon Simple Storage Service (Amazon S3) console.

In this example, you create a `FindMatches` transform to find matching records, teach it how to identify matching and nonmatching records, and use it in an AWS Glue job. The AWS Glue job writes a new Amazon S3 file with an additional column named `match_id`.

The source data used by this tutorial is a file named `dblp_acm_records.csv`. This file is a modified version of academic publications (DBLP and ACM) available from the original DBLP ACM dataset. The `dblp_acm_records.csv` file is a comma-separated values (CSV) file in UTF-8 format with no byte-order mark (BOM).

A second file, `dblp_acm_labels.csv`, is an example labeling file that contains both matching and nonmatching records used to teach the transform as part of the tutorial.

Topics

- Step 1: Crawl the source data
- Step 2: Add a machine learning transform
- Step 3: Teach your machine learning transform
- Step 4: Estimate the quality of your machine learning transform
- Step 5: Add and run a job with your machine learning transform
- Step 6: Verify output data from Amazon S3

Step 1: Crawl the source data

First, crawl the source Amazon S3 CSV file to create a corresponding metadata table in the Data Catalog.

> **Important**
>
> To direct the crawler to create a table for only the CSV file, store the CSV source data in a different Amazon S3 folder from other files.

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, choose **Crawlers**, **Add crawler**.
3. Follow the wizard to create and run a crawler named `demo-crawl-dblp-acm` with output to database `demo-db-dblp-acm`. When running the wizard, create the database `demo-db-dblp-acm` if it doesn't already exist. Choose an Amazon S3 include path to sample data in the current AWS Region. For example, for `us-east-1`, the Amazon S3 include path to the source file is `s3://ml-transforms-public-datasets-us-east-1/dblp-acm/records/dblp_acm_records.csv`.

   If successful, the crawler creates the table `dblp_acm_records_csv` with the following columns: id, title, authors, venue, year, and source.

Step 2: Add a machine learning transform

Next, add a machine learning transform that is based on the schema of your data source table created by the crawler named `demo-crawl-dblp-acm`.

1. On the AWS Glue console, in the navigation pane under **Data Integration and ETL**, choose **Data classification tools > Record Matching**, then **Add transform**. Follow the wizard to create a `Find matches` transform with the following properties.
   a. For **Transform name**, enter **`demo-xform-dblp-acm`**. This is the name of the transform that is used to find matches in the source data.
   b. For **IAM role**, choose an IAM role that has permission to the Amazon S3 source data, labeling file, and AWS Glue API operations. For more information, see [Create an IAM Role for AWS Glue](https://docs.aws.amazon.com/glue/) in the *AWS Glue Developer Guide*.
   c. For **Data source**, choose the table named **dblp_acm_records_csv** in database **demo-db-dblp-acm**.
   d. For **Primary key**, choose the primary key column for the table, **id**.

2. In the wizard, choose **Finish** and return to the **ML transforms** list.

Step 3: Teach your machine learning transform

Next, you teach your machine learning transform using the tutorial sample labeling file.

You can't use a machine language transform in an extract, transform, and load (ETL) job until its status is **Ready for use**. To get your transform ready, you must teach it how to identify matching and nonmatching records by providing examples of matching and nonmatching records. To teach your transform, you can **Generate a label file**, add labels, and then **Upload label file**. In this tutorial, you can use the example labeling file named `dblp_acm_labels.csv`. For more information about the labeling process, see Labeling.

1. On the AWS Glue console, in the navigation pane, choose **Record Matching**.
2. Choose the `demo-xform-dblp-acm` transform, and then choose **Action**, **Teach**. Follow the wizard to teach your `Find matches` transform.
3. On the transform properties page, choose **I have labels**. Choose an Amazon S3 path to the sample labeling file in the current AWS Region. For example, for `us-east-1`, upload the provided labeling file from the Amazon S3 path `s3://ml-transforms-public-datasets-us-east-1/dblp-acm/labels/dblp_acm_labels.csv` with the option to **overwrite** existing labels. The labeling file must be located in Amazon S3 in the same Region as the AWS Glue console.

   When you upload a labeling file, a task is started in AWS Glue to add or overwrite the labels used to teach the transform how to process the data source.
4. On the final page of the wizard, choose **Finish**, and return to the **ML transforms** list.

Step 4: Estimate the quality of your machine learning transform

Next, you can estimate the quality of your machine learning transform. The quality depends on how much labeling you have done. For more information about estimating quality, see Estimate quality.

1. On the AWS Glue console, in the navigation pane under **Data Integration and ETL**, choose **Data classification tools > Record Matching**.
2. Choose the `demo-xform-dblp-acm` transform, and choose the **Estimate quality** tab. This tab displays the current quality estimates, if available, for the transform.

3. Choose **Estimate quality** to start a task to estimate the quality of the transform. The accuracy of the quality estimate is based on the labeling of the source data.
4. Navigate to the **History** tab. In this pane, task runs are listed for the transform, including the **Estimating quality** task. For more details about the run, choose **Logs**. Check that the run status is **Succeeded** when it finishes.

Step 5: Add and run a job with your machine learning transform

In this step, you use your machine learning transform to add and run a job in AWS Glue. When the transform `demo-xform-dblp-acm` is **Ready for use**, you can use it in an ETL job.

1. On the AWS Glue console, in the navigation pane, choose **Jobs**.
2. Choose **Add job**, and follow the steps in the wizard to create an ETL Spark job with a generated script. Choose the following property values for your transform:
   a. For **Name**, choose the example job in this tutorial, **demo-etl-dblp-acm**.
   b. For **IAM role**, choose an IAM role with permission to the Amazon S3 source data, labeling file, and AWS Glue API operations. For more information, see Create an IAM Role for AWS Glue in the *AWS Glue Developer Guide*.
   c. For **ETL language**, choose **Scala**. This is the programming language in the ETL script.
   d. For **Script file name**, choose **demo-etl-dblp-acm**. This is the file name of the Scala script (same as the job name).
   e. For **Data source**, choose **dblp_acm_records_csv**. The data source you choose must match the machine learning transform data source schema.
   f. For **Transform type**, choose **Find matching records** to create a job using a machine learning transform.
   g. Clear **Remove duplicate records**. You don't want to remove duplicate records because the output records written have an additional `match_id` field added.
   h. For **Transform**, choose **demo-xform-dblp-acm**, the machine learning transform used by the job.
   i. For **Create tables in your data target**, choose to create tables with the following properties:
      - **Data store type** — **`Amazon S3`**
      - **Format** — **CSV**

  - **Compression type — None**
  - **Target path —** The Amazon S3 path where the output of the job is written (in the current console AWS Region)

3. Choose **Save job and edit script** to display the script editor page.
4. Edit the script to add a statement to cause the job output to the **Target path** to be written to a single partition file. Add this statement immediately following the statement that runs the `FindMatches` transform. The statement is similar to the following.

   ```
   val single_partition = findmatches1.repartition(1)
   ```

   You must modify the `.writeDynamicFrame(findmatches1)` statement to write the output as `.writeDynamicFrame(single_partion)`.
5. After you edit the script, choose **Save**. The modified script looks similar to the following code, but customized for your environment.

   ```

   import com.amazonaws.services.glue.GlueContext
   import com.amazonaws.services.glue.errors.CallSite
   import com.amazonaws.services.glue.ml.FindMatches
   import com.amazonaws.services.glue.util.GlueArgParser
   import com.amazonaws.services.glue.util.Job
   import com.amazonaws.services.glue.util.JsonOptions
   import org.apache.spark.SparkContext
   import scala.collection.JavaConverters._

   object GlueApp {
     def main(sysArgs: Array[String]) {
       val spark: SparkContext = new SparkContext()
       val glueContext: GlueContext = new GlueContext(spark)
       // @params: [JOB_NAME]
       val args = GlueArgParser.getResolvedOptions(sysArgs, Seq("JOB_NAME").toArray)
       Job.init(args("JOB_NAME"), glueContext, args.asJava)
       // @type: DataSource
       // @args: [database = "demo-db-dblp-acm", table_name = "dblp_acm_records_csv",
    transformation_ctx = "datasource0"]
       // @return: datasource0
       // @inputs: []
       val datasource0 = glueContext.getCatalogSource(database = "demo-db-dblp-acm",
    tableName = "dblp_acm_records_csv", redshiftTmpDir = "", transformationContext =
    "datasource0").getDynamicFrame()
   ```

```
    // @type: FindMatches
    // @args: [transformId = "tfm-123456789012", emitFusion = false,
 survivorComparisonField = "<primary_id>", transformation_ctx = "findmatches1"]
    // @return: findmatches1
    // @inputs: [frame = datasource0]
    val findmatches1 = FindMatches.apply(frame = datasource0, transformId
 = "tfm-123456789012", transformationContext = "findmatches1",
 computeMatchConfidenceScores = true)


    // Repartition the previous DynamicFrame into a single partition.
    val single_partition = findmatches1.repartition(1)


    // @type: DataSink
    // @args: [connection_type = "s3", connection_options = {"path": "s3://aws-
glue-ml-transforms-data/sal"}, format = "csv", transformation_ctx = "datasink2"]
    // @return: datasink2
    // @inputs: [frame = findmatches1]
    val datasink2 = glueContext.getSinkWithFormat(connectionType =
 "s3", options = JsonOptions("""{"path": "s3://aws-glue-ml-transforms-
data/sal"}"""), transformationContext = "datasink2", format =
 "csv").writeDynamicFrame(single_partition)
    Job.commit()
  }
}
```

6. Choose **Run job** to start the job run. Check the status of the job in the jobs list. When the job finishes, in the **ML transform**, **History** tab, there is a new **Run ID** row added of type **ETL job**.
7. Navigate to the **Jobs**, **History** tab. In this pane, job runs are listed. For more details about the run, choose **Logs**. Check that the run status is **Succeeded** when it finishes.

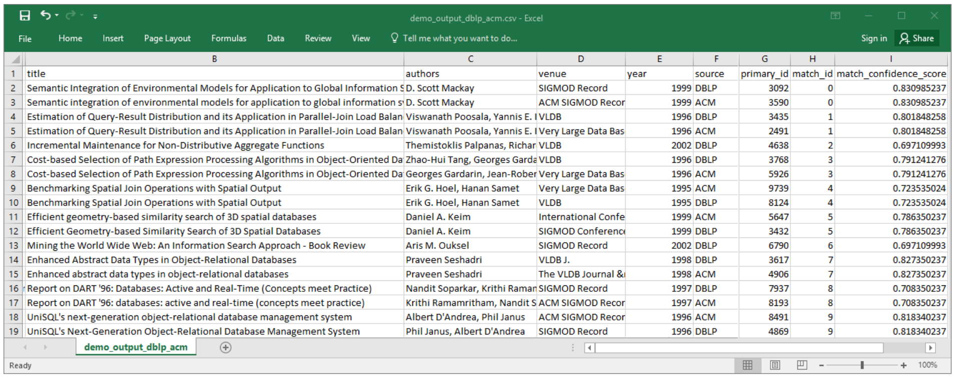
Step 6: Verify output data from Amazon S3

In this step, you check the output of the job run in the Amazon S3 bucket that you chose when you added the job. You can download the output file to your local machine and verify that matching records were identified.

1. Open the Amazon S3 console at https://console.aws.amazon.com/s3/.

2. Download the target output file of the job `demo-etl-dblp-acm`. Open the file in a spreadsheet application (you might need to add a file extension `.csv` for the file to properly open).

   The following image shows an excerpt of the output in Microsoft Excel.



	B	C	D	E	F	G	H	I
1	title	authors	venue	year	source	primary_id	match_id	match_confidence_score
2	Semantic Integration of Environmental Models for Application to Global Information S	D. Scott Mackay	SIGMOD Record	1999	DBLP	3092	0	0.830985237
3	Semantic integration of environmental models for application to global information sy	D. Scott Mackay	ACM SIGMOD Recor	1999	ACM	3590	0	0.830985237
4	Estimation of Query-Result Distribution and its Application in Parallel-Join Load Balan	Viswanath Poosala, Yannis E. I	VLDB	1996	DBLP	3435	1	0.801848258
5	Estimation of Query-Result Distribution and its Application in Parallel-Join Load Balan	Viswanath Poosala, Yannis E. I	Very Large Data Bas	1996	ACM	2491	1	0.801848258
6	Incremental Maintenance for Non-Distributive Aggregate Functions	Themistoklis Palpanas, Richar	VLDB	2002	DBLP	4638	2	0.697109993
7	Cost-based Selection of Path Expression Processing Algorithms in Object-Oriented Da	Zhao-Hui Tang, Georges Garda	VLDB	1996	DBLP	3768	3	0.791241276
8	Cost-based Selection of Path Expression Processing Algorithms in Object-Oriented Da	Georges Gardarin, Jean-Rober	Very Large Data Bas	1996	ACM	5926	3	0.791241276
9	Benchmarking Spatial Join Operations with Spatial Output	Erik G. Hoel, Hanan Samet	Very Large Data Bas	1995	ACM	9739	4	0.723535024
10	Benchmarking Spatial Join Operations with Spatial Output	Erik G. Hoel, Hanan Samet	VLDB	1995	DBLP	8124	4	0.723535024
11	Efficient geometry-based similarity search of 3D spatial databases	Daniel A. Keim	International Confe	1999	ACM	5647	5	0.786350237
12	Efficient Geometry-based Similarity Search of 3D Spatial Databases	Daniel A. Keim	SIGMOD Conference	1999	DBLP	3432	5	0.786350237
13	Mining the World Wide Web: An Information Search Approach - Book Review	Aris M. Ouksel	SIGMOD Record	2002	DBLP	6790	6	0.697109993
14	Enhanced Abstract Data Types in Object-Relational Databases	Praveen Seshadri	VLDB J.	1998	DBLP	3617	7	0.827350237
15	Enhanced abstract data types in object-relational databases	Praveen Seshadri	The VLDB Journal &	1998	ACM	4906	7	0.827350237
16	Report on DART '96: Databases: Active and Real-Time (Concepts meet Practice)	Nandit Soparkar, Krithi Raman	SIGMOD Record	1997	DBLP	7937	8	0.708350237
17	Report on DART '96: databases: active and real-time (concepts meet practice)	Krithi Ramamritham, Nandit S	ACM SIGMOD Recor	1997	ACM	8193	8	0.708350237
18	UniSQL's next-generation object-relational database management system	Albert D'Andrea, Phil Janus	ACM SIGMOD Recor	1996	ACM	8491	9	0.818340237
19	UniSQL's Next-Generation Object-Relational Database Management System	Phil Janus, Albert D'Andrea	SIGMOD Record	1996	DBLP	4869	9	0.818340237

   The data source and target file both have 4,911 records. However, the `Find matches` transform adds another column named `match_id` to identify matching records in the output. Rows with the same `match_id` are considered matching records. The `match_confidence_score` is a number between 0 and 1 that provides an estimate of the quality of matches found by `Find matches`.

3. Sort the output file by `match_id` to easily see which records are matches. Compare the values in the other columns to see if you agree with the results of the `Find matches` transform. If you don't, you can continue to teach the transform by adding more labels.

   You can also sort the file by another field, such as `title`, to see if records with similar titles have the same `match_id`.

Finding incremental matches

The Find matches feature allows you to identify duplicate or matching records in your dataset, even when the records don’t have a common unique identifier and no fields match exactly. The initial release of the Find matches transform identified matching records within a single dataset. When you add new data to the dataset, you had to merge it with the existing clean dataset and rerun matching against the complete merged dataset.

The incremental matching feature makes it simpler to match to incremental records against existing matched datasets. Suppose that you want to match prospects data with existing customer datasets. The incremental match capability provides you the flexibility to match hundreds of thousands of new prospects with an existing database of prospects and customers by merging the results into a single database or table. By matching only between the new and existing datasets, the find incremental matches optimization reduces computation time, which also reduces cost.

The usage of incremental matching is similar to Find matches as described in Tutorial: Creating a machine learning transform with AWS Glue. This topic identifies only the differences with incremental matching.

For more information, see the blog post on Incremental data matching.

Running an incremental matching job

For the following procedure, suppose the following:

- You have crawled the existing dataset into the table *first_records*. The *first_records* dataset must be a matched dataset, or the output of the matched job.
- You have created and trained a Find matches transform with AWS Glue version 2.0. This is the only version of AWS Glue that supports incremental matches.
- The ETL language is Scala. Note that Python is also supported.
- The model already generated is called `demo-xform`.

1. Crawl the incremental dataset to the table *second_records*.
2. On the AWS Glue console, in the navigation pane, choose **Jobs**.
3. Choose **Add job**, and follow the steps in the wizard to create an ETL Spark job with a generated script. Choose the following property values for your transform:
    a. For **Name**, choose **demo-etl**.
    b. For **IAM role**, choose an IAM role with permission to the Amazon S3 source data, labeling file, and AWS Glue API operations.
    c. For **ETL language**, choose **Scala**.
    d. For **Script file name**, choose **demo-etl**. This is the file name of the Scala script.
    e. For **Data source**, choose **first_records**. The data source you choose must match the machine learning transform data source schema.
    f. For **Transform type**, choose **Find matching records** to create a job using a machine learning transform.

   g. Select the incremental matching option, and for **Data Source** select the table named **second_records**.

   h. For **Transform**, choose **demo-xform**, the machine learning transform used by the job.

   i. Choose **Create tables in your data target** or **Use tables in the data catalog and update your data target**.

4. Choose **Save job and edit script** to display the script editor page.

5. Choose **Run job** to start the job run.

Using FindMatches in a visual job

To use the **FindMatches** transform in AWS Glue Studio, you can use the **Custom Transform** node that invokes the FindMatches API. For more information on how to use a custom transform, see Creating a custom transformation

> **Note**
>
> Currently, the FindMatches API only works with `Glue 2.0`. In order to run a job with the Custom transform that invokes the FindMatches API, ensure the AWS Glue version is `Glue 2.0` in the **Job details** tab. If the version of AWS Glue is not `Glue 2.0`, the job will fail at runtime with the following error message: “cannot import name 'FindMatches' from 'awsglueml.transforms'”.

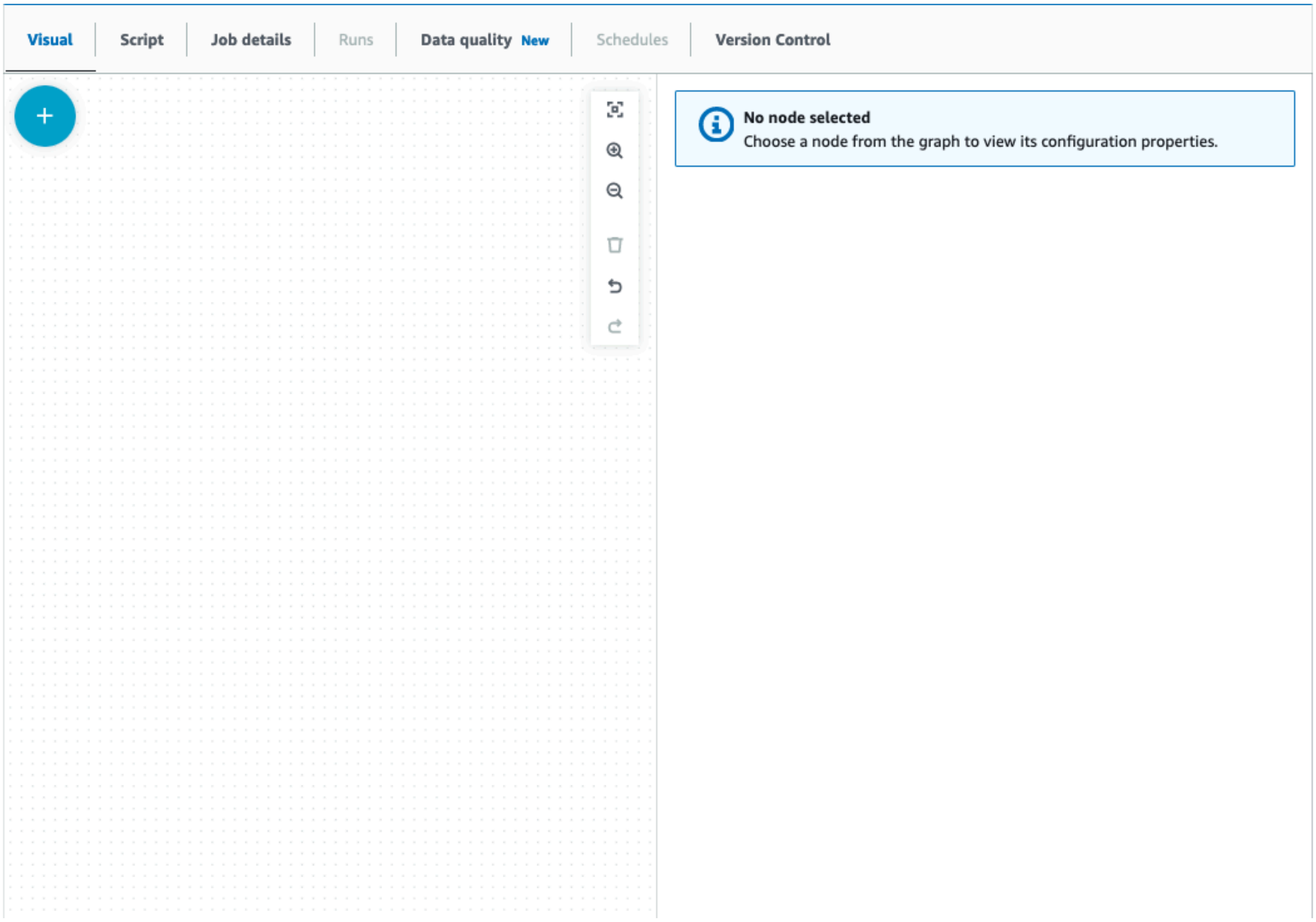
Prerequisites

- In order to use the **Find Matches** transform, open the AWS Glue Studio console at https://console.aws.amazon.com/gluestudio/.
- Create a machine learning transform. When created, a transformId is generated. You will need this ID for the steps below. For more information on how to create a machine learning transform, see Adding and editing machine learning transforms.

Adding a FindMatches transform

To add a FindMatches transform:

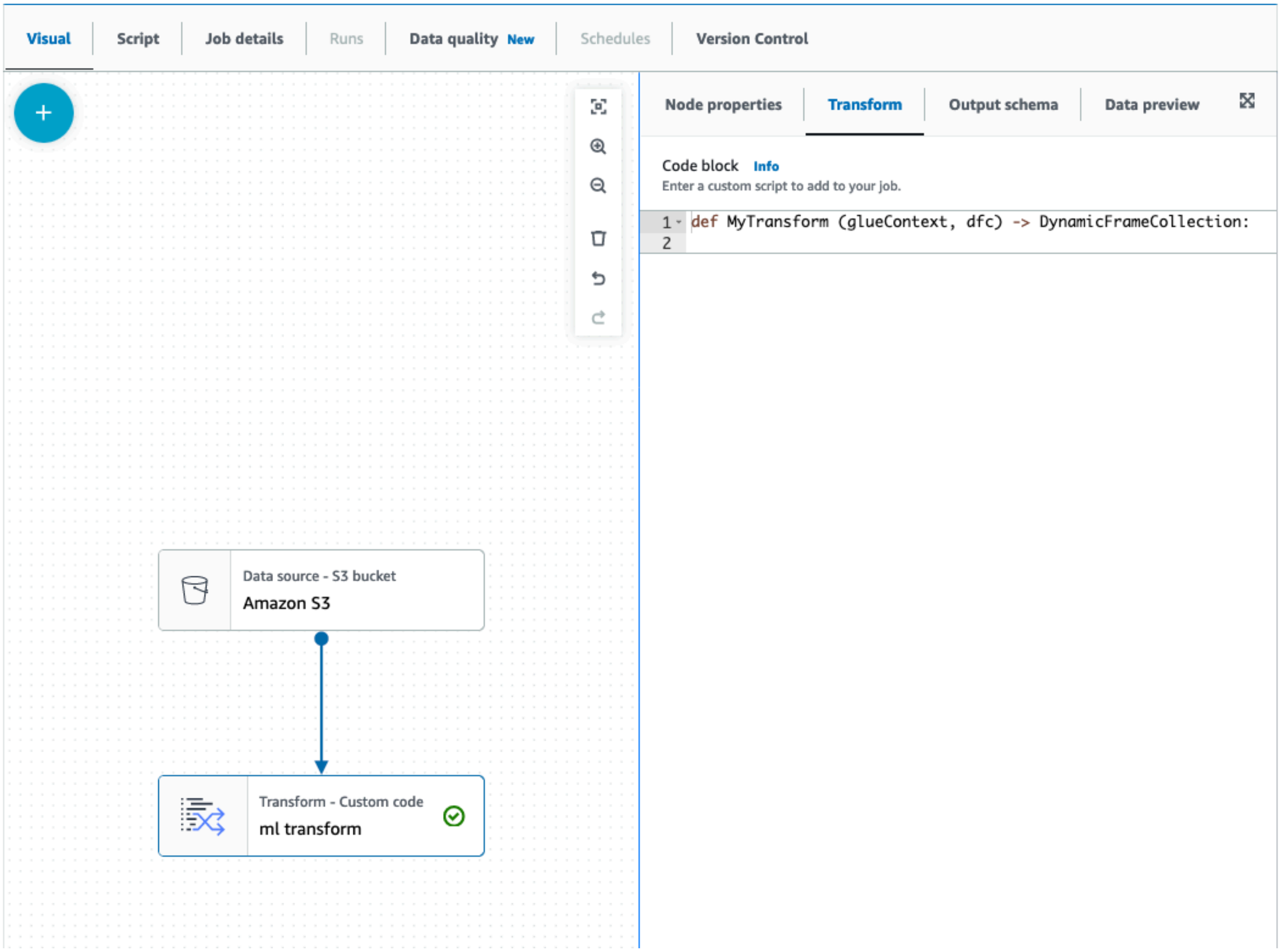
1. In the AWS Glue Studio job editor, open the Resource panel by clicking on the cross symbol in the upper left-hand corner of the visual job graph and choose a Data source by choosing the **Data tab**. This is the data source you want to check for matches.



No node selected
Choose a node from the graph to view its configuration properties.

2. Choose the data source node, then open the Resource panel by clicking on the cross symbol in the upper left-hand corner of the visual job graph and search for 'custom transform'. Choose the **Custom Transform** node to add it to the graph. The **Custom Transform** is linked to the data source node. If it is not, you can click on the **Custom Transform** node and choose the **Node properties** tab, then under **Node parents**, choose the data source.
3. Click the **Custom Transform** node in the visual graph, then choose the **Node properties** tab and name the custom transform. It is recommended that you rename the transform so that the transform name is easily identifiable in the visual graph.

4. Choose the **Transform** tab, where you can edit the code block. This is where the code to invoke the FindMatches API can be added.


Visual
Script
Job details
Runs
Data quality New
Schedules
Version Control
Node properties
Transform
Output schema
Data preview
Data source - S3 bucket
Amazon S3
Transform - Custom code
ml transform

Code block Info

Enter a custom script to add to your job.

```
def MyTransform (glueContext, dfc) -> DynamicFrameCollection:
```

The code block contains pre-populated code to get you started. Overwrite the pre-populated code with the template below. The template has a placeholder for the **transformId**, which you can provide.

```
def MyTransform (glueContext, dfc) -> DynamicFrameCollection:
    dynf = dfc.select(list(dfc.keys())[0])
    from awsglueml.transforms import FindMatches
    findmatches = FindMatches.apply(frame = dynf, transformId = "<your id>")
    return(DynamicFrameCollection({"FindMatches": findmatches}, glueContext))
```

5. Click the **Custom Transform** node in the visual graph, then open the Resource panel by clicking on the cross symbol in the upper left-hand corner of the visual job graph and search

for 'Select From Collection'. There is no need to change the default selection since there is only one DynamicFrame in the collection.

6. You can continue adding transformations or store the result, which is now enriched with the find matches additional columns. If you want to reference those new columns in downstream transforms, you need to add them to the transform output schema. the easiest way to do that is to choose the **Data preview** tab and then in the schema tab choose “Use datapreview schema”.
7. To customize FindMatches, you can add additional parameters to pass to the 'apply' method. See FindMatches class.

Adding a FindMatches incrementally transformation

In the case of incremental matches, the process is the same as **Adding a FindMatches transformation** with the following differences:

- Instead of a parent node for the custom transform, you need two parent nodes.
- The first parent node should be the dataset.
- The second parent node should be the incremental dataset.

  Replace the `transformId` with your `transformId` in the template code block:

```
def MyTransform (glueContext, dfc) -> DynamicFrameCollection:
    dfs = list(dfc.values())
    dynf = dfs[0]
    inc_dynf = dfs[1]
    from awsglueml.transforms import FindIncrementalMatches
    findmatches = FindIncrementalMatches.apply(existingFrame = dynf, incrementalFrame
 = inc_dynf,
                                    transformId = "<your id>")
    return(DynamicFrameCollection({"FindMatches": findmatches}, glueContext))
```

- For optional parameters, see FindIncrementalMatches class.

Migrate Apache Spark programs to AWS Glue

Apache Spark is an open-source platform for distributed computing workloads performed on large datasets. AWS Glue leverages Spark's capabilities to provide an optimized experience for ETL. You can migrate Spark programs to AWS Glue to take advantage of our features. AWS Glue provides the same performance enhancements you would expect from Apache Spark on Amazon EMR.

Run Spark code

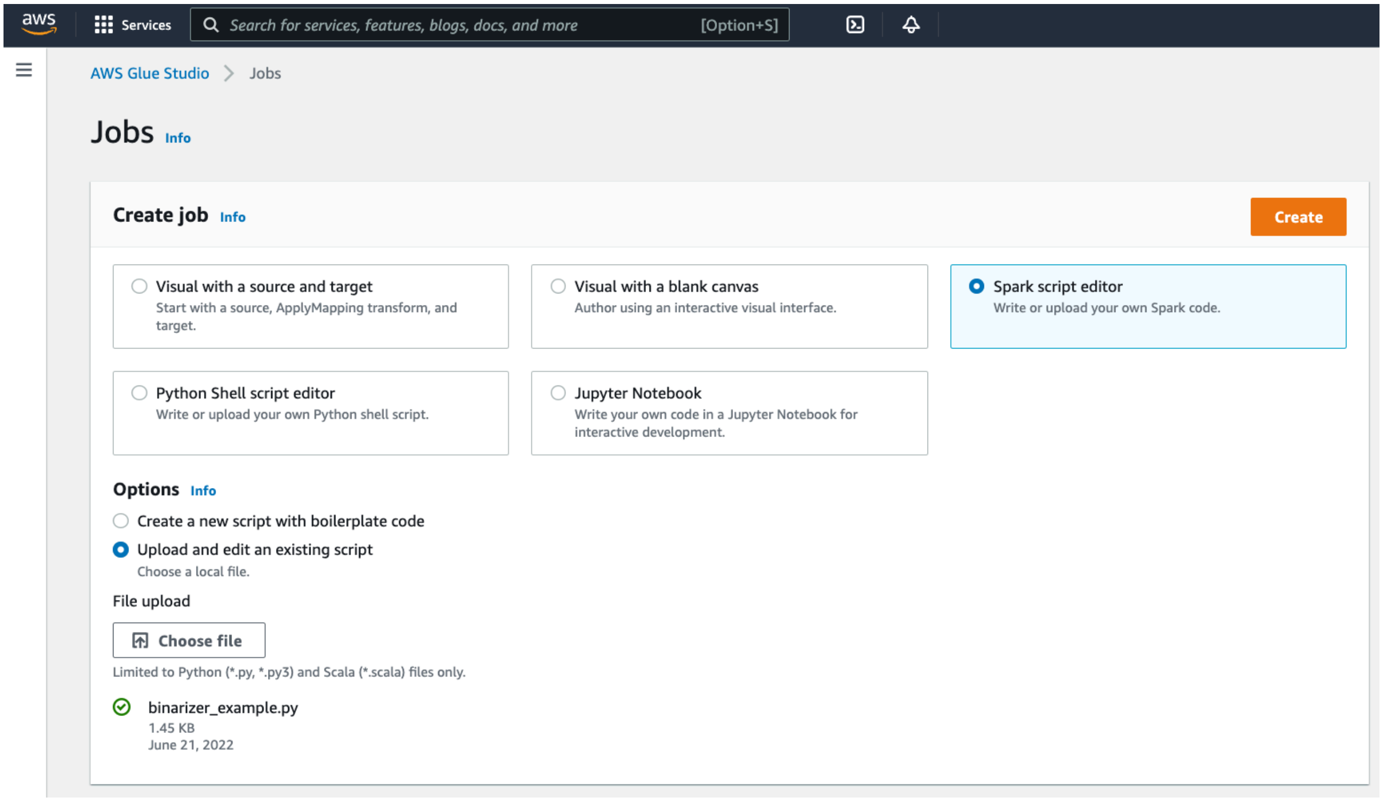
Native Spark code can be run in a AWS Glue environment out of the box. Scripts are often developed by iteratively changing a piece of code, a workflow suited for an Interactive Session. However, existing code is more suited to run in a AWS Glue job, which allows you to schedule and consistently get logs and metrics for each script run. You can upload and edit an existing script through the console.

1. Acquire the source to your script. For this example, you will use an example script from the Apache Spark repository. Binarizer Example
2. In the AWS Glue Console, expand the left-side navigation pane and select **ETL** > **Jobs**

   In the **Create job** panel, select **Spark script editor**. An **Options** section will appear. Under **Options**, select **Upload and edit an existing script**.

   A **File upload** section will appear. Under **File upload**, click **Choose file**. Your system file chooser will appear. Navigate to the location where you saved `binarizer_example.py`, select it and confirm your selection.

   A **Create** button will appear on the header for the **Create job** panel. Click it.


aws
Services
Search for services, features, blogs, docs, and more [Option+S]
AWS Glue Studio > Jobs
Create

Jobs Info

Create job Info

○ Visual with a source and target
Start with a source, ApplyMapping transform, and target.

○ Visual with a blank canvas
Author using an interactive visual interface.

● Spark script editor
Write or upload your own Spark code.

○ Python Shell script editor
Write or upload your own Python shell script.

○ Jupyter Notebook
Write your own code in a Jupyter Notebook for interactive development.

Options Info

○ Create a new script with boilerplate code

● Upload and edit an existing script
Choose a local file.

File upload

Choose file

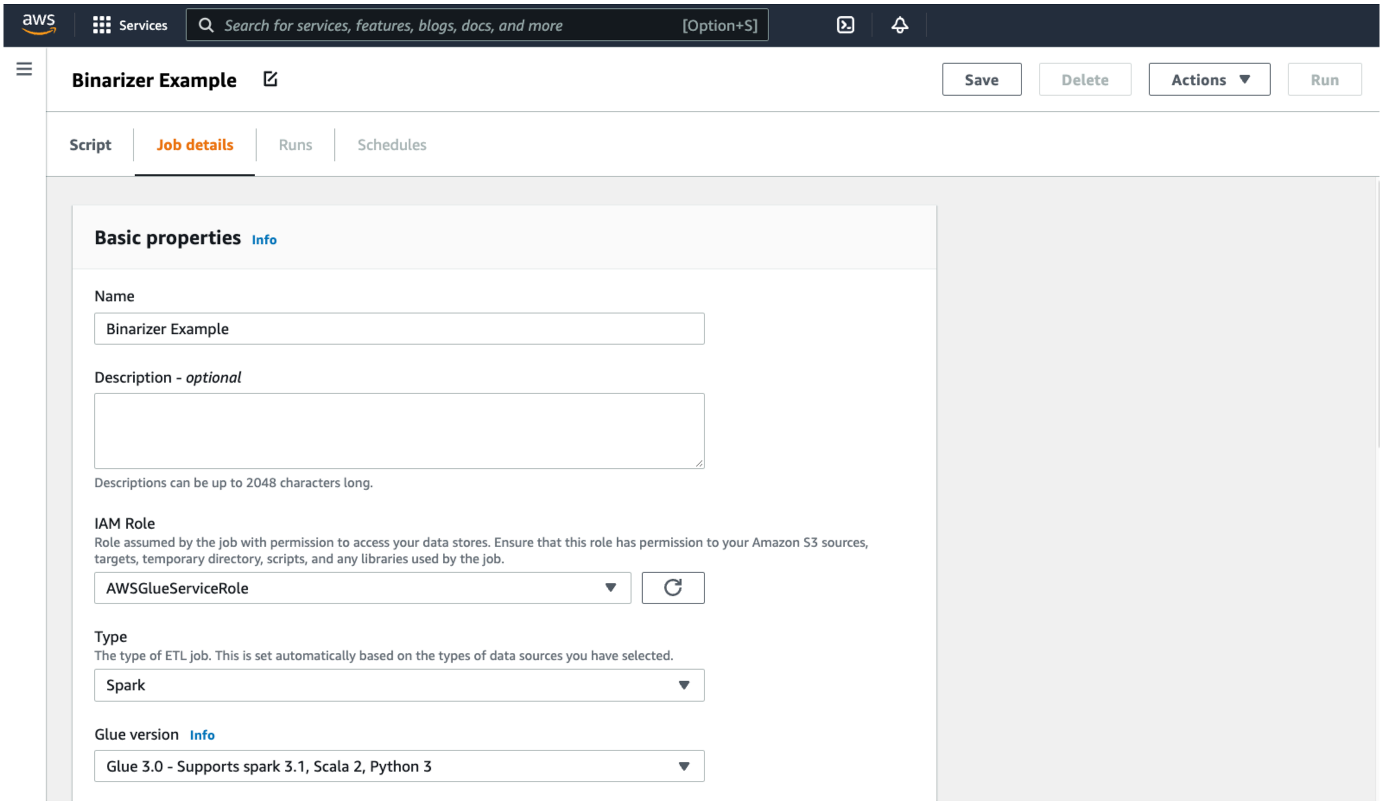
Limited to Python (*.py, *.py3) and Scala (*.scala) files only.

✓ binarizer_example.py
1.45 KB
June 21, 2022

3. Your browser will navigate to the script editor. On the header, click the **Job details** tab. Set the **Name** and **IAM Role**. For guidance around AWS Glue IAM roles, consult the section called "Setting up IAM permissions".

   Optionally - set **Requested number of workers** to 2 and **Number of retries** to 1. These options are valuable when running production jobs, but turning them down will streamline your experience while testing out a feature.

   In the title bar, click **Save**, then **Run**


aws
Services
Search for services, features, blogs, docs, and more
[Option+S]
Save
Delete
Actions
Run
Script
Job details
Runs
Schedules
Info
Info

Binarizer Example

Basic properties

Name

Binarizer Example

Description - *optional*

Descriptions can be up to 2048 characters long.

IAM Role

Role assumed by the job with permission to access your data stores. Ensure that this role has permission to your Amazon S3 sources, targets, temporary directory, scripts, and any libraries used by the job.

AWSGlueServiceRole

Type

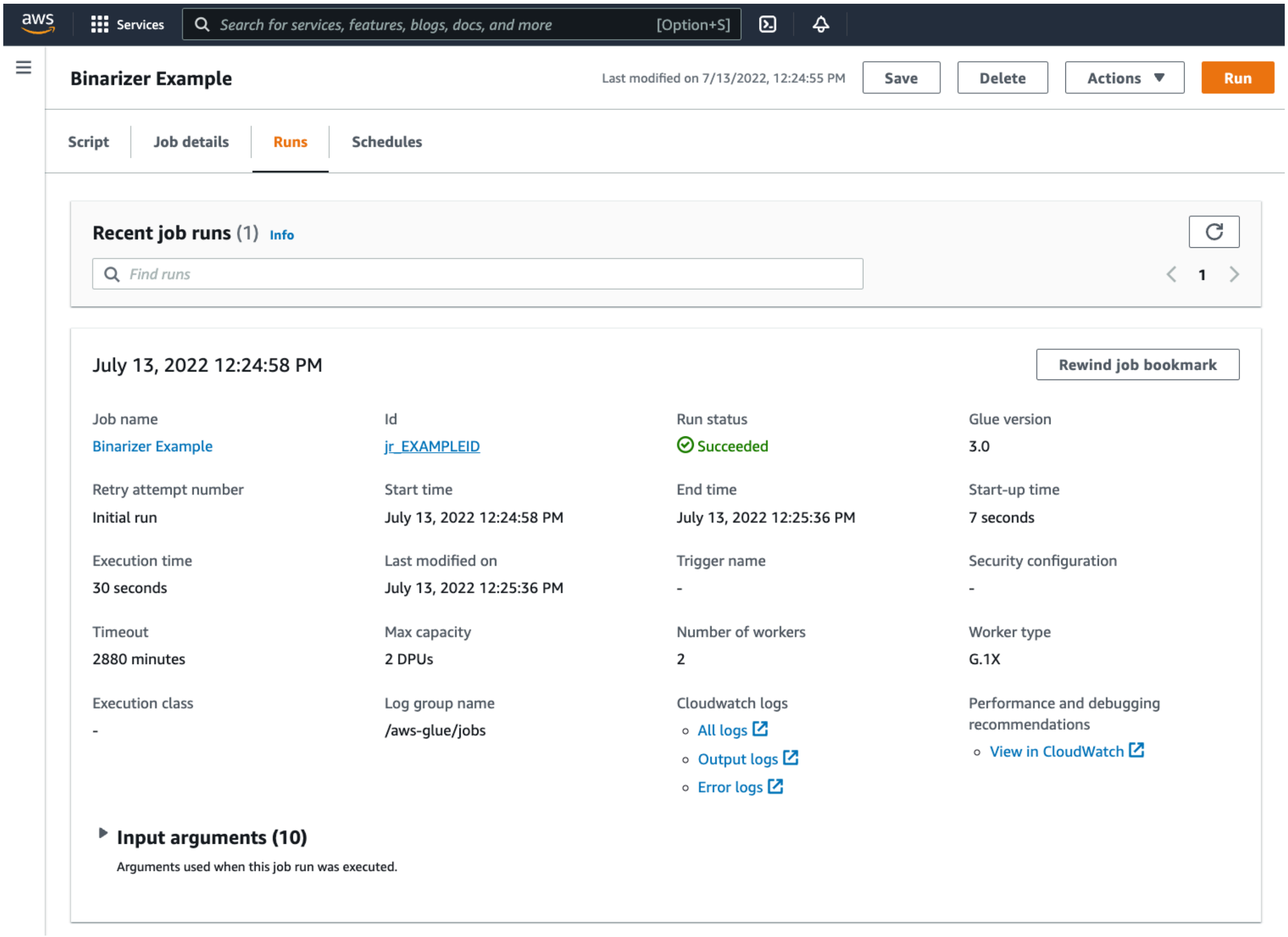
The type of ETL job. This is set automatically based on the types of data sources you have selected.

Spark

Glue version

Glue 3.0 - Supports spark 3.1, Scala 2, Python 3

4. Navigate to the **Runs** tab. You will see a panel corresponding to your job run. Wait a few minutes and the page should automatically refresh to show **Succeeded** under **Run status**.



Binarizer Example

Last modified on 7/13/2022, 12:24:55 PM

Recent job runs (1)

July 13, 2022 12:24:58 PM

Job name: Binarizer Example

Id: jr_EXAMPLEID

Run status: Succeeded

Glue version: 3.0

Retry attempt number: Initial run

Start time: July 13, 2022 12:24:58 PM

End time: July 13, 2022 12:25:36 PM

Start-up time: 7 seconds

Execution time: 30 seconds

Last modified on: July 13, 2022 12:25:36 PM

Trigger name: -

Security configuration: -

Timeout: 2880 minutes

Max capacity: 2 DPUs

Number of workers: 2

Worker type: G.1X

Execution class: -

Log group name: /aws-glue/jobs

Cloudwatch logs:
- All logs
- Output logs
- Error logs

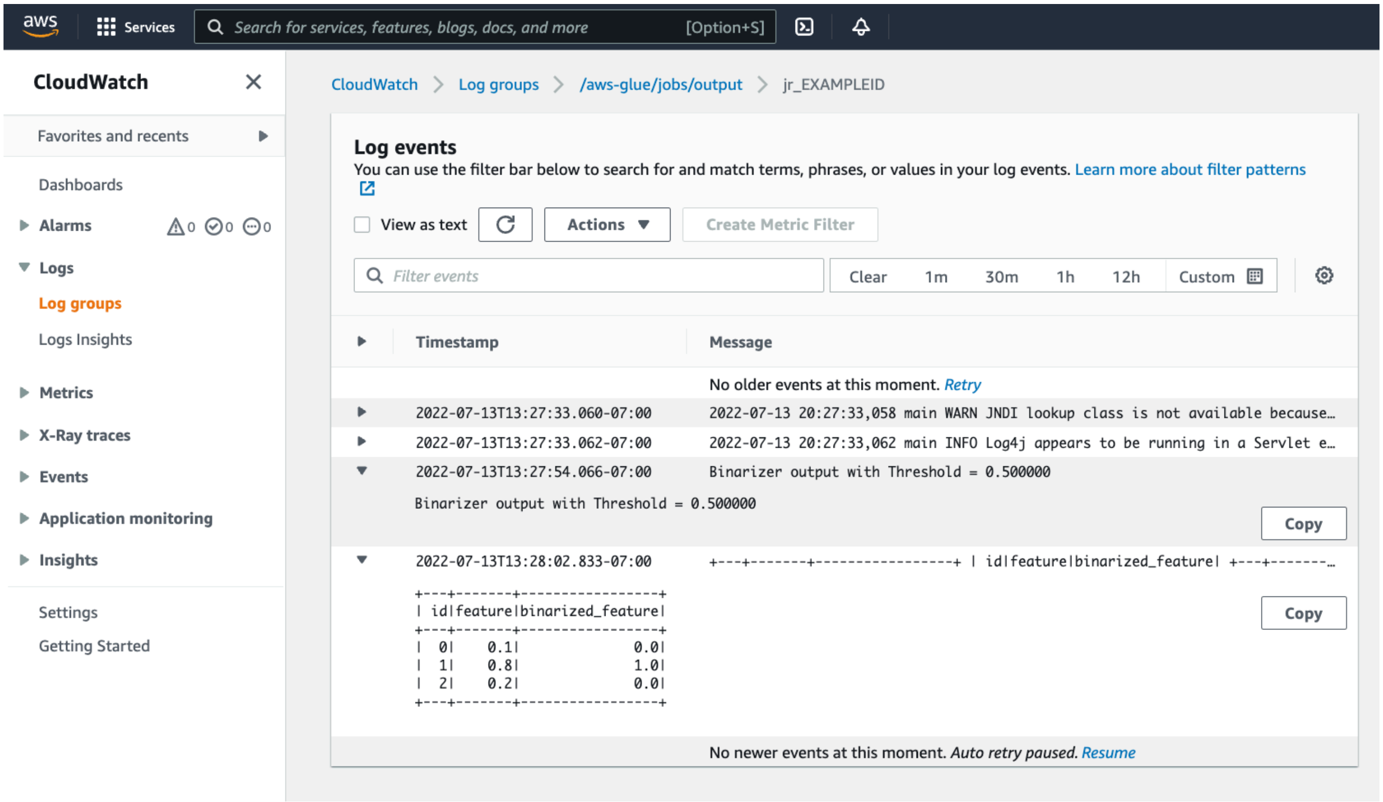
Performance and debugging recommendations:
- View in CloudWatch

Input arguments (10)

Arguments used when this job run was executed.

5. You will want to examine your output to confirm that the Spark script ran as intended. This Apache Spark sample script should write a string to the output stream. You can find that by navigating to **Output logs** under **Cloudwatch logs** in the panel for the successful job run. Note the job run id, a generated id under the **Id** label beginning with `jr_`.

   This will open the CloudWatch console, set to visualize the contents of the default AWS Glue log group `/aws-glue/jobs/output`, filtered to the contents of the log streams for the job run id. Each worker will have generated a log stream, shown as rows under the **Log streams** . One worker should have run the requested code. You will need to open all the log streams to identify the correct worker. Once you find the right worker, you should see the output of the script, as seen in the following image:


aws
Services
Search for services, features, blogs, docs, and more
[Option+S]
CloudWatch
Favorites and recents
Dashboards
Alarms
0
0
0
Logs
Log groups
Logs Insights
Metrics
X-Ray traces
Events
Application monitoring
Insights
Settings
Getting Started
CloudWatch > Log groups > /aws-glue/jobs/output > jr_EXAMPLEID
☐ View as text
Actions
Create Metric Filter
Filter events
Clear
1m
30m
1h
12h
Custom
Timestamp
Message
Copy
Copy

Log events

You can use the filter bar below to search for and match terms, phrases, or values in your log events. Learn more about filter patterns

No older events at this moment. *Retry*

```
2022-07-13T13:27:33.060-07:00   2022-07-13 20:27:33,058 main WARN JNDI lookup class is not available because…
2022-07-13T13:27:33.062-07:00   2022-07-13 20:27:33,062 main INFO Log4j appears to be running in a Servlet e…
2022-07-13T13:27:54.066-07:00   Binarizer output with Threshold = 0.500000
Binarizer output with Threshold = 0.500000
2022-07-13T13:28:02.833-07:00   +---+-------+-----------------+ | id|feature|binarized_feature| +---+-------+…
+---+-------+-----------------+
| id|feature|binarized_feature|
+---+-------+-----------------+
|  0|    0.1|              0.0|
|  1|    0.8|              1.0|
|  2|    0.2|              0.0|
+---+-------+-----------------+
```

No newer events at this moment. *Auto retry paused.* Resume

Common procedures needed for migrating Spark programs

Assess Spark version support

AWS Glue release versions define the version of Apache Spark and Python available to the AWS Glue job. You can find our AWS Glue versions and what they support at the section called “AWS Glue versions”. You may need to update your Spark program to be compatible with a newer version of Spark in order to access certain AWS Glue features.

Include third-party libraries

Many existing Spark programs will have dependencies, both on private and public artifacts. AWS Glue supports JAR style dependencies for Scala Jobs as well as Wheel and source pure-Python dependencies for Python jobs.

Python - For information about Python dependencies, see the section called “Python libraries”

Common Python dependencies are provided in the AWS Glue environment, including the commonly requested Pandas library. Dependencies are included in AWS Glue Version 2.0+. For more information about provided modules, see the section called “Python modules already provided in AWS Glue”. If you need to supply a Job with a different version of a dependency included by default, you can use `--additional-python-modules`. For information about job arguments, see the section called “Job parameters”.

You can supply additional Python dependencies with the `--extra-py-files` job argument. If you are migrating a job from a Spark program, this parameter is a good option because it is functionally equivalent to the `--py-files` flag in PySpark, and is subject to the same limitations. For more information about the `--extra-py-files` parameter, see the section called “Including Python files with PySpark native features”

For new jobs, you can manage Python dependencies with the `--additional-python-modules` job argument. Using this argument allows for a more thorough dependency management experience. This parameter supports Wheel style dependencies, including those with native code bindings compatible with Amazon Linux 2.

Scala

You can supply additional Scala dependencies with the `--extra-jars` Job Argument. Dependencies must be hosted in Amazon S3 and the argument value should be a comma delimited list of Amazon S3 paths with no spaces. You may find it easier to manage your configuration by rebundling your dependencies before hosting and configuring them. AWS Glue JAR dependencies contain Java bytecode, which can be generated from any JVM language. You can use other JVM languages, such as Java, to write custom dependencies.

Manage data source credentials

Existing Spark programs may come with complex or custom configuration to pull data from their datasources. Common datasource auth flows are supported by AWS Glue connections. For more information about AWS Glue connections, see Connecting to data.

AWS Glue connections facilitate connecting your Job to a variety of types of data stores in two primary ways: through method calls to our libraries and setting the **Additional network connection** in the AWS console. You may also call the AWS SDK from within your job to retrieve information from a connection.

Method calls – AWS Glue Connections are tightly integrated with the AWS Glue Data Catalog, a service that allows you to curate information about your datasets, and the methods available to interact with AWS Glue connections reflect that. If you have an existing auth configuration you would like to reuse, for JDBC connections, you can access your AWS Glue connection configuration through the `extract_jdbc_conf` method on the `GlueContext`. For more information, see the section called “extract_jdbc_conf”

Console configuration – AWS Glue Jobs use associated AWS Glue connections to configure connections to Amazon VPC subnets. If you directly manage your security materials, you may need to provide a `NETWORK` type **Additional network connection** in the AWS console to configure routing. For more information about the AWS Glue connection API, see the section called "Connections"

If your Spark programs has a custom or uncommon auth flow, you may need to manage your security materials in a hands-on fashion. If AWS Glue connections do not seem like a good fit, you can securely host security materials in Secrets Manager and access them through the boto3 or AWS SDK, which are provided in the job.

Configure Apache Spark

Complex migrations often alter Spark configuration to acommodate their workloads. Modern versions of Apache Spark allow runtime configuration to be set with the `SparkSession`. AWS Glue 3.0+ Jobs are provided a `SparkSession`, which can be modified to set runtime configuration. Apache Spark Configuration. Tuning Spark is complex, and AWS Glue does not guarantee support for setting all Spark configuration. if your migration requires substantial Spark-level configuration, contact support.

Set custom configuration

Migrated Spark programs may be designed to take custom configuration. AWS Glue allows configuration to be set on the job and job run level, through the job arguments. For information about job arguments, see the section called "Job parameters". You can access job arguments within the context of a job through our libraries. AWS Glue provides a utility function to provide a consistent view between arguments set on the job and arguments set on the job run. See the section called "getResolvedOptions" in Python and the section called "GlueArgParser" in Scala.

Migrate Java code

As explained in the section called "Third-party libraries", your dependencies can contain classes generated by JVM languages, such as Java or Scala. Your dependencies can include a `main` method. You can use a `main` method in a dependency as the entrypoint for a AWS Glue Scala job. This allows you to write your `main` method in Java, or reuse a `main` method packaged to your own library standards.

To use a `main` method from a dependency, perform the following: Clear the contents of the editing pane providing the default `GlueApp` object. Provide the fully qualified name of a class in a dependency as a job argument with the key `--class`. You should then be able to trigger a Job run.

You cannot configure the order or structure of the arguments AWS Glue passes to the `main` method. If your existing code needs to read configuration set in AWS Glue, this will likely cause incompatibility with prior code. If you use `getResolvedOptions`, you will also not have a good place to call this method. Consider invoking your dependency directly from a main method generated by AWS Glue. The following AWS Glue ETL script shows an example of this.

```
import com.amazonaws.services.glue.util.GlueArgParser

object GlueApp {
  def main(sysArgs: Array[String]) {
    val args = GlueArgParser.getResolvedOptions(sysArgs, Seq("JOB_NAME").toArray)

    // Invoke static method from JAR. Pass some sample arguments as a String[], one
 defined inline and one taken from the job arguments, using getResolvedOptions
    com.mycompany.myproject.MyClass.myStaticPublicMethod(Array("string parameter1",
 args("JOB_NAME")))

    // Alternatively, invoke a non-static public method.
    (new com.mycompany.myproject.MyClass).someMethod()
  }
}
```

Working with Ray jobs in AWS Glue

This section provides information about using AWS Glue for Ray jobs. For more information about writing AWS Glue for Ray scripts, consult the the section called “AWS Glue for Ray” section.

Topics

- Getting started with AWS Glue for Ray
- Supported Ray runtime environments
- Accounting for workers in Ray jobs
- Using job parameters in Ray jobs
- Monitoring Ray jobs with metrics

Getting started with AWS Glue for Ray

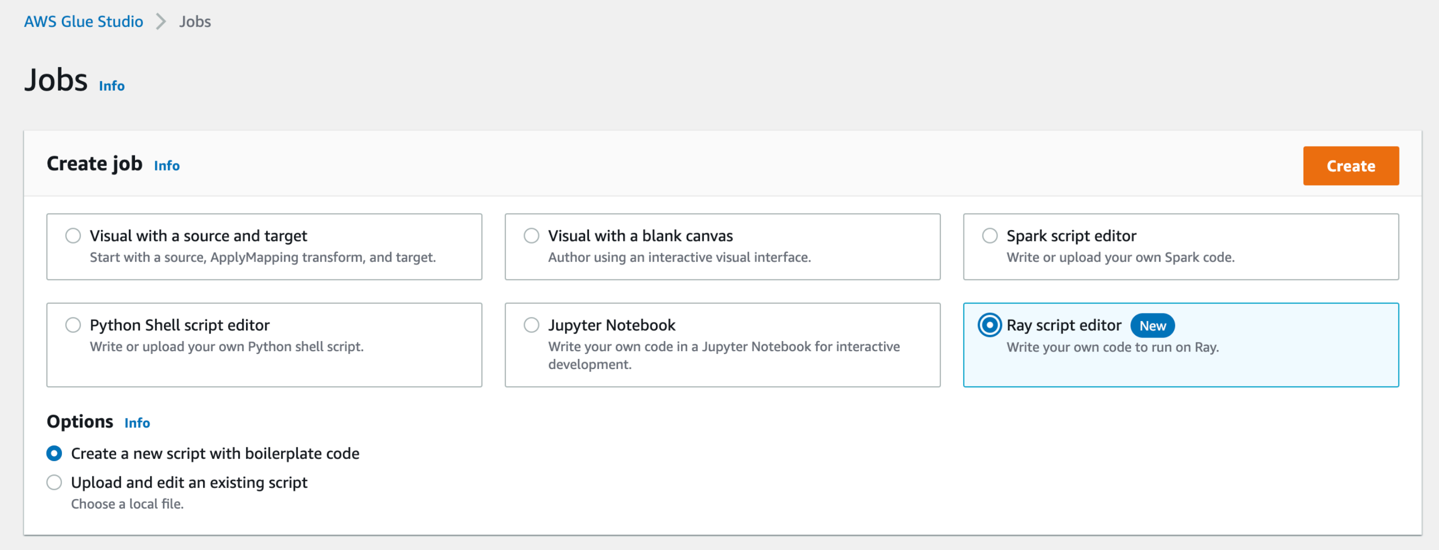
To work with AWS Glue for Ray, you use the same AWS Glue jobs and interactive sessions that you use with AWS Glue for Spark. AWS Glue jobs are designed for running the same script on a recurring cadence, while interactive sessions are designed to let you run snippets of code sequentially against the same provisioned resources.

AWS Glue ETL and Ray are different underneath, so in your script, you have access to different tools, features, and configuration. As a new computation framework managed by AWS Glue, Ray has a different architecture and uses different vocabulary to describe what it does. For more information, see Architecture Whitepapers in the Ray documentation.

> **Note**
>
> AWS Glue for Ray is available in US East (N. Virginia), US East (Ohio), US West (Oregon), Asia Pacific (Tokyo), and Europe (Ireland).

Ray jobs in the AWS Glue Studio console

On the **Jobs** page in the AWS Glue Studio console, you can select a new option when you're creating a job in AWS Glue Studio—**Ray script editor**. Choose this option to create a Ray job in the console. For more information about jobs and how they're used, see *Building visual ETL jobs*.



AWS Glue Studio > Jobs

Jobs Info

Create job Info

Create

○ Visual with a source and target
Start with a source, ApplyMapping transform, and target.

○ Visual with a blank canvas
Author using an interactive visual interface.

○ Spark script editor
Write or upload your own Spark code.

○ Python Shell script editor
Write or upload your own Python shell script.

○ Jupyter Notebook
Write your own code in a Jupyter Notebook for interactive development.

● Ray script editor New
Write your own code to run on Ray.

Options Info

● Create a new script with boilerplate code

○ Upload and edit an existing script
Choose a local file.

Ray jobs in the AWS CLI and SDK

Ray jobs in the AWS CLI use the same SDK actions and parameters as other jobs. AWS Glue for Ray introduces new values for certain parameters. For more information in the Jobs API, see the section called “Jobs”.

Supported Ray runtime environments

In Spark jobs, `GlueVersion` determines the versions of Apache Spark and Python available in an AWS Glue for Spark job. The Python version indicates the version that is supported for jobs of type Spark. This is not how Ray runtime environments are configured.

For Ray jobs, you should set `GlueVersion` to `4.0` or greater. However, the versions of Ray, Python, and additional libraries that are available in your Ray job are determined by the `Runtime` field in the job definition.

The `Ray2.4` runtime environment will be available for a minimum of 6 months after release. As Ray rapidly evolves, you will be able to incorporate Ray updates and improvements through future runtime environment releases.

Valid values: `Ray2.4`

Runtime value	Ray and Python versions
`Ray2.4` (for AWS Glue 4.0+)	Ray 2.4.0 Python 3.9

Additional information

- For release notes that accompany AWS Glue on Ray releases, see the section called “AWS Glue versions”.
- For Python libraries that are provided in a runtime environment, see the section called “Modules provided with Ray jobs”.

Accounting for workers in Ray jobs

AWS Glue runs Ray jobs on new Graviton-based EC2 worker types, which are only available for Ray jobs. To appropriately provision these workers for the workloads Ray is designed for, we provide a different ratio of compute resources to memory resources from most workers. In order to account for these resources, we use the memory-optimized data processing unit (M-DPU) rather than the standard data processing unit (DPU).

- One M-DPU corresponds to 4 vCPUs and 32 GB of memory.

- One DPU corresponds to 4 vCPUs and 16 GB of memory. DPUs are used to account for resources in AWS Glue with Spark jobs and corresponding workers.

Ray jobs currently have access to one worker type, `Z.2X`. The `Z.2X` worker maps to 2 M-DPUs (8 vCPUs, 64 GB of memory) and has 128 GB of disk space. A `Z.2X` machine provides 8 Ray workers (one per vCPU).

The number of M-DPUs that you can use concurrently in an account is subject to a service quota. For more information about your AWS Glue account limits, see AWS Glue endpoints and quotas.

You specify the number of worker nodes that are available to a Ray job with `--number-of-workers (NumberOfWorkers)` in the job definition. For more information about Ray values in the Jobs API, see the section called "Jobs".

You can further specify a minimum number of workers that a Ray job must allocate with the `--min-workers` job parameter. For more information about job parameters, see the section called "Reference".

Using job parameters in Ray jobs

You set arguments for AWS Glue Ray jobs the same way you set arguments for AWS Glue for Spark jobs. For more information about the AWS Glue API, see the section called "Jobs". You can configure AWS Glue Ray jobs with different arguments, which are listed in this reference. You can also provide your own arguments.

You can configure a job through the console, on the **Job details** tab, under the **Job Parameters** heading. You can also configure a job through the AWS CLI by setting `DefaultArguments` on a job, or setting `Arguments` on a job run. Default arguments and job parameters stay with the job through multiple runs.

For example, the following is the syntax for running a job using `--arguments` to set a special parameter.

```
$ aws glue start-job-run --job-name "CSV to CSV" --arguments='--scriptLocation="s3://
my_glue/libraries/test_lib.py",--test-environment="true"'
```

After you set the arguments, you can access job parameters from within your Ray job through environment variables. This gives you a way to configure your job for each run. The name of the environment variable will be the job argument name without the `--` prefix.

For instance, in the previous example, the variable names would be `scriptLocation` and `test-environment`. You would then retrieve the argument through methods available in the standard library: `test_environment = os.environ.get('test-environment')`. For more information about accessing environment variables with Python, see os module in the Python documentation.

Configure how Ray jobs generate logs

By default, Ray jobs generate logs and metrics that are sent to CloudWatch and Amazon S3. You can use the `--logging_configuration` parameter to alter how logs are generated, currently you can use it to stop Ray jobs from generating various types of logs. This parameter takes a JSON object, whose keys correspond to the logs/behaviors you would like to alter. It supports the following keys:

- `CLOUDWATCH_METRICS` – Configures CloudWatch metrics series that can be used to visualize job health. For more information about metrics, see the section called “Ray job metrics”.
- `CLOUDWATCH_LOGS` – Configures CloudWatch logs that provide Ray application level details about the status the job run. For more information about logs, see the section called “Troubleshooting Ray errors”.
- `S3` – Configures what AWS Glue writes to Amazon S3, primarily similar information to CloudWatch logs but as files rather than log streams.

To disable a Ray logging behavior, provide the value `{\"IS_ENABLED\": \"False\"}`. For example, to disable CloudWatch metrics and CloudWatch logs, provide the following configuration:

```
"--logging_configuration": "{\"CLOUDWATCH_METRICS\": {\"IS_ENABLED\": \"False\"},
 \"CLOUDWATCH_LOGS\": {\"IS_ENABLED\": \"False\"}}"
```

Reference

Ray jobs recognize the following argument names that you can use to set up the script environment for your Ray jobs and job runs:

- `--logging_configuration` – Used to stop the generation of various logs created by Ray jobs. These logs are generated by default on all Ray jobs. Format: String-escaped JSON object. For more information, see the section called “Configure how Ray jobs generate logs”.
- `--min-workers` – The minimum number of worker nodes that are allocated to a Ray job. A worker node can run multiple replicas, one per virtual CPU. Format: integer. Minimum: 0.

Maximum: value specified in `--number-of-workers (NumberOfWorkers)` on the job definition. For more information about accounting for worker nodes, see the section called “Accounting for workers in Ray jobs”.

- `--object_spilling_config` – AWS Glue for Ray supports using Amazon S3 as a way of extending the space available to Ray's object store. To enable this behavior, you can provide Ray an *object spilling*JSON config object with this parameter. For more information about Ray object spilling configuration, see Object Spilling in the Ray documentation. Format: JSON object.

  AWS Glue for Ray only supports spilling to disk or spilling to Amazon S3 at once. You can provide multiple locations for spilling, as long as they respect this limitation. When spilling to Amazon S3, you will also need to add IAM permissions to your job for this bucket.

  When providing a JSON object as configuration with the CLI, you must provide it as a string, with the JSON object string-escaped. For example, a string value for spilling to one Amazon S3 path would look like: `"{\"type\": \"smart_open\", \"params\": {\"uri\": \"s3path\"}}"`. In AWS Glue Studio, provide this parameter as a JSON object with no extra formatting.
- `--object_store_memory_head` – The memory allocated to the Plasma object store on the Ray head node. This instance runs cluster management services, as well as worker replicas. The value represents a percentage of free memory on the instance after a warm start. You use this parameter to tune memory intensive workloads—defaults are acceptable for most use cases. Format: positive integer. Minimum: 1. Maximum: 100.

  For more information about Plasma, see The Plasma In-Memory Object Store in the Ray documentation.
- `--object_store_memory_worker` – The memory allocated to the Plasma object store on the Ray worker nodes. These instances only run worker replicas. The value represents a percentage of free memory on the instance after a warm start. This parameter is used to tune memory intensive workloads—defaults are acceptable for most use cases. Format: positive integer. Minimum: 1. Maximum: 100.

  For more information about Plasma, see The Plasma In-Memory Object Store in the Ray documentation.
- `--pip-install` – A set of Python packages to be installed. You can install packages from PyPI using this argument. Format: comma-delimited list.

A PyPI package entry is in the format `package==version`, with the PyPI name and version of your target package. Entries use Python version matching to match the package and version, such as ==, not the single equals =. There are other version-matching operators. For more information, see PEP 440 on the Python website. You can also provide custom modules with `--s3-py-modules`.

- `--s3-py-modules` – A set of Amazon S3 paths that host Python module distributions. Format: comma-delimited list.

  You can use this to distribute your own modules to your Ray job. You can also provide modules from PyPI with `--pip-install`. Unlike with AWS Glue ETL, custom modules are not set up through pip, but are passed to Ray for distribution. For more information, see the section called "Additional Python modules for Ray jobs".
- `--working-dir` – A path to a .zip file hosted in Amazon S3 that contains files to be distributed to all nodes running your Ray job. Format: string. For more information, see the section called "Providing files to your Ray job".

Monitoring Ray jobs with metrics

You can monitor Ray jobs using AWS Glue Studio and Amazon CloudWatch. CloudWatch collects and processes raw metrics from AWS Glue with Ray, which makes them available for analysis. These metrics are visualized in the AWS Glue Studio console, so you can monitor your job as it runs.

For a general overview of how to monitor AWS Glue, see the section called "Using CloudWatch metrics". For a general overview of how to use CloudWatch metrics that are published by AWS Glue, see the section called "Monitoring with CloudWatch".

Monitoring Ray jobs in the AWS Glue console

On the details page for a job run, below the **Run details** section, you can view pre-built aggregated graphs that visualize your available job metrics. AWS Glue Studio sends job metrics to CloudWatch for every job run. With these, you can build a profile of your cluster and tasks, as well as access detailed information about each node.

For more information about available metrics graphs, see the section called "Viewing Amazon CloudWatch metrics for a Ray job run".

Overview of Ray jobs metrics in CloudWatch

We publish Ray metrics when detailed monitoring is enabled in CloudWatch. Metrics are published to the `Glue/Ray` CloudWatch namespace.

- **Instance metrics**

  We publish metrics about the CPU, memory and disk utilization of instances assigned to a job. These metrics are identified by features such as `ExecutorId`, `ExecutorType` and `host`. These metrics are a subset of the standard Linux CloudWatch agent metrics. You can find information about metric names and features in the CloudWatch documentation. For more information, see Metrics collected by the CloudWatch agent.
- **Ray cluster metrics**

  We forward metrics from the Ray processes that run your script to this namespace, then provide those most critical for you. The metrics that are available might differ by Ray version. For more information about which Ray version your job is running, see the section called "AWS Glue versions".

  Ray collects metrics at the instance level. It also provides metrics for tasks and the cluster. For more information about Ray's underlying metric strategy, see Metrics in the Ray documentation.

> **Note**
>
> We don't publish Ray metrics to the `Glue/Job Metrics/` namespace, which is only used for AWS Glue ETL jobs.

Configuring job properties for Python shell jobs in AWS Glue

You can use a Python shell job to run Python scripts as a shell in AWS Glue. With a Python shell job, you can run scripts that are compatible with Python 3.6 or Python 3.9.

Note

Support for Pyshell v3.6 will end on March 1, 2026. To migrate your workloads, see Migrate from AWS Glue Python shell jobs. If you wish to continue with Python shell v3.9 see Migrating from Python shell 3.6 to Python shell 3.9.

Topics

- Limitations
- Defining job properties for Python shell jobs
- Supported libraries for Python shell jobs
- Providing your own Python library
- Use AWS CloudFormation with Python shell jobs in AWS Glue
- Migrating from Python shell 3.6 to Python shell 3.9
- Migrate from AWS Glue Python shell jobs

Limitations

Note the following limitations of Python Shell jobs:

- You can't use job bookmarks with Python shell jobs.
- You can't package any Python libraries as `.egg` files in Python 3.9+. Instead, use `.whl`.
- The `--extra-files` option cannot be used, because of a limitation on temporary copies of S3 data.

Defining job properties for Python shell jobs

These sections describe defining job properties in AWS Glue Studio, or using the AWS CLI.

AWS Glue Studio

When you define your Python shell job in AWS Glue Studio, you provide some of the following properties:

IAM role

Specify the AWS Identity and Access Management (IAM) role that is used for authorization to resources that are used to run the job and access data stores. For more information about permissions for running jobs in AWS Glue, see Identity and access management for AWS Glue.

Type

Choose **Python shell** to run a Python script with the job command named `pythonshell`.

Python version

Choose the Python version. The default is Python 3.9. Valid versions are Python 3.6 and Python 3.9.

Load common analytics libraries (Recommended)

Choose this option to include common libraries for Python 3.9 in the Python shell.

If your libraries are either custom or they conflict with the pre-installed ones, you can choose not to install common libraries. However, you can install additional libraries besides the common libraries.

When you select this option, the `library-set` option is set to `analytics`. When you de-select this option, the `library-set` option is set to none.

Script filename and Script path

The code in the script defines your job's procedural logic. You provide the script name and location in Amazon Simple Storage Service (Amazon S3). Confirm that there isn't a file with the same name as the script directory in the path. To learn more about using scripts, see AWS Glue programming guide.

Script

The code in the script defines your job's procedural logic. You can code the script in Python 3.6 or Python 3.9. You can edit a script in AWS Glue Studio.

Data processing units

The maximum number of AWS Glue data processing units (DPUs) that can be allocated when this job runs. A DPU is a relative measure of processing power that consists of 4 vCPUs of compute capacity and 16 GB of memory. For more information, see AWS Glue pricing.

You can set the value to 0.0625 or 1. The default is 0.0625. In either case, the local disk for the instance will be 20GB.

CLI

You can also create a **Python shell** job using the AWS CLI, as in the following example.

```
aws glue create-job --name python-job-cli --role Glue_DefaultRole
    --command '{"Name" :  "pythonshell", "PythonVersion": "3.9", "ScriptLocation" :
"s3://amzn-s3-demo-bucket/scriptname.py"}'
    --max-capacity 0.0625
```

Note

You don't need to specify the version of AWS Glue since the parameter `--glue-version` doesn't apply for AWS Glue shell jobs. Any version specified will be ignored.

Jobs that you create with the AWS CLI default to Python 3. Valid Python versions are 3 (corresponding to 3.6), and 3.9. To specify Python 3.6, add this tuple to the `--command` parameter: `"PythonVersion":"3"`

To specify Python 3.9, add this tuple to the `--command` parameter: `"PythonVersion":"3.9"`

To set the maximum capacity used by a Python shell job, provide the `--max-capacity` parameter. For Python shell jobs, the `--allocated-capacity` parameter can't be used.

Supported libraries for Python shell jobs

In Python shell using Python 3.9, you can choose the library set to use pre-packaged library sets for your needs. You can use the `library-set` option to choose the library set. Valid values are `analytics`, and none.

The environment for running a Python shell job supports the following libraries:

Python version	**Python 3.6**	**Python 3.9**	
Library set	**N/A**	**analytics**	**none**
avro		1.11.0	
awscli	116.242	1.23.5	1.23.5

Python version	Python 3.6	Python 3.9	
awswrangler		2.15.1	
botocore	1.12.232	1.24.21	1.23.5
boto3	1.9.203	1.21.21	
elasticsearch		8.2.0	
numpy	1.16.2	1.22.3	
pandas	0.24.2	1.4.2	
psycopg2		2.9.3	
pyathena		2.5.3	
PyGreSQL	5.0.6		
PyMySQL		1.0.2	
pyodbc		4.0.32	
pyorc		0.6.0	
redshift-connector		2.0.907	
requests	2.22.0	2.27.1	
scikit-learn	0.20.3	1.0.2	
scipy	1.2.1	1.8.0	
SQLAlchemy		1.4.36	
s3fs		2022.3.0	

You can use the `NumPy` library in a Python shell job for scientific computing. For more information, see NumPy. The following example shows a NumPy script that can be used in a Python shell job. The script prints "Hello world" and the results of several mathematical calculations.

```
import numpy as np
print("Hello world")

a = np.array([20,30,40,50])
print(a)

b = np.arange( 4 )

print(b)

c = a-b

print(c)

d = b**2

print(d)
```

Providing your own Python library

Using PIP

Python shell using Python 3.9 lets you provide additional Python modules or different versions at the job level. You can use the `--additional-python-modules` option with a list of comma-separated Python modules to add a new module or change the version of an existing module. You cannot provide custom Python modules hosted on Amazon S3 with this parameter when using Python shell jobs.

For example to update or to add a new `scikit-learn` module use the following key and value: `"--additional-python-modules", "scikit-learn==0.21.3"`.

AWS Glue uses the Python Package Installer (pip3) to install the additional modules. You can pass additional pip3 options inside the `--additional-python-modules` value. For example, `"scikit-learn==0.21.3 -i https://pypi.python.org/simple/"`. Any incompatibilities or limitations from pip3 apply.

Note

To avoid incompatibilities in the future, we recommend that you use libraries built for Python 3.9.

Using an Egg or Whl file

You might already have one or more Python libraries packaged as an `.egg` or a `.whl` file. If so, you can specify them to your job using the AWS Command Line Interface (AWS CLI) under the "`--extra-py-files`" flag, as in the following example.

```
aws glue create-job --name python-redshift-test-cli --role role --command '{"Name" :
 "pythonshell", "ScriptLocation" : "s3://MyBucket/python/library/redshift_test.py"}'
     --connections Connections=connection-name --default-arguments '{"--extra-py-
files" : ["s3://amzn-s3-demo-bucket/EGG-FILE", "s3://amzn-s3-demo-bucket/WHEEL-FILE"]}'
```

If you aren't sure how to create an `.egg` or a `.whl` file from a Python library, use the following steps. This example is applicable on macOS, Linux, and Windows Subsystem for Linux (WSL).

To create a Python .egg or .whl file

1. Create an Amazon Redshift cluster in a virtual private cloud (VPC), and add some data to a table.
2. Create an AWS Glue connection for the VPC-SecurityGroup-Subnet combination that you used to create the cluster. Test that the connection is successful.
3. Create a directory named `redshift_example`, and create a file named `setup.py`. Paste the following code into `setup.py`.

```
from setuptools import setup

setup(
    name="redshift_module",
    version="0.1",
    packages=['redshift_module']
)
```

4. In the `redshift_example` directory, create a `redshift_module` directory. In the `redshift_module` directory, create the files `__init__.py` and `pygresql_redshift_common.py`.
5. Leave the `__init__.py` file empty. In `pygresql_redshift_common.py`, paste the following code. Replace *`port`*, *`db_name`*, *`user`*, and *`password_for_user`* with details specific to your Amazon Redshift cluster. Replace *`table_name`* with the name of the table in Amazon Redshift.

```
import pg


def get_connection(host):
    rs_conn_string = "host=%s port=%s dbname=%s user=%s password=%s" % (
        host, port, db_name, user, password_for_user)

    rs_conn = pg.connect(dbname=rs_conn_string)
    rs_conn.query("set statement_timeout = 1200000")
    return rs_conn


def query(con):
    statement = "Select * from table_name;"
    res = con.query(statement)
    return res
```

6. If you're not already there, change to the `redshift_example` directory.
7. Do one of the following:

   - To create an `.egg` file, run the following command.

     ```
     python setup.py bdist_egg
     ```

   - To create a `.whl` file, run the following command.

     ```
     python setup.py bdist_wheel
     ```

8. Install the dependencies that are required for the preceding command.
9. The command creates a file in the `dist` directory:

   - If you created an egg file, it's named `redshift_module-0.1-py2.7.egg`.

- If you created a wheel file, it's named `redshift_module-0.1-py2.7-none-any.whl`.

  Upload this file to Amazon S3.

  In this example, the uploaded file path is either *`s3://amzn-s3-demo-bucket/EGG-FILE`* or *`s3://amzn-s3-demo-bucket/WHEEL-FILE`*.

10. Create a Python file to be used as a script for the AWS Glue job, and add the following code to the file.

```
from redshift_module import pygresql_redshift_common as rs_common

con1 = rs_common.get_connection(redshift_endpoint)
res = rs_common.query(con1)

print "Rows in the table cities are: "

print res
```

11. Upload the preceding file to Amazon S3. In this example, the uploaded file path is *`s3://amzn-s3-demo-bucket/scriptname.py`*.

12. Create a Python shell job using this script. On the AWS Glue console, on the **Job properties** page, specify the path to the `.egg/.whl` file in the **Python library path** box. If you have multiple `.egg/.whl` files and Python files, provide a comma-separated list in this box.

    When modifying or renaming `.egg` files, the file names must use the default names generated by the "python setup.py bdist_egg" command or must adhere to the Python module naming conventions. For more information, see the Style Guide for Python Code.

    Using the AWS CLI, create a job with a command, as in the following example.

```
aws glue create-job --name python-redshift-test-cli --role Role --command
 '{"Name" :  "pythonshell", "ScriptLocation" : "s3://amzn-s3-demo-bucket/
scriptname.py"}'
     --connections Connections="connection-name" --default-arguments '{"--extra-
py-files" : ["s3://amzn-s3-demo-bucket/EGG-FILE", "s3://amzn-s3-demo-bucket/WHEEL-
FILE"]}'
```

When the job runs, the script prints the rows created in the *`table_name`* table in the Amazon Redshift cluster.

Use AWS CloudFormation with Python shell jobs in AWS Glue

You can use AWS CloudFormation with Python shell jobs in AWS Glue. The following is an example:

```
AWSTemplateFormatVersion: 2010-09-09
Resources:
  Python39Job:
    Type: 'AWS::Glue::Job'
    Properties:
      Command:
        Name: pythonshell
        PythonVersion: '3.9'
        ScriptLocation: 's3://bucket/location'
      MaxRetries: 0
      Name: python-39-job
      Role: RoleName
```

The Amazon CloudWatch Logs group for Python shell jobs output is `/aws-glue/python-jobs/output`. For errors, see the log group `/aws-glue/python-jobs/error`.

Migrating from Python shell 3.6 to Python shell 3.9

To migrate your Python shell jobs to the latest AWS Glue version:

1. In the AWS Glue console (https://console.aws.amazon.com/glue/), choose your existing Python shell job.
2. In the **Job** details tab, set the Python version to `Python 3.9` and choose **Save**.
3. Ensure that your job script is compatible with Python 3.9 and that it runs successfully.

Migrate from AWS Glue Python shell jobs

AWS launched the AWS Glue Python shell jobs in 2018 AWS launched the AWS Glue Python shell jobs in 2018 in order to give customers an easy way to run Python scripts for small-to-medium sized ETL jobs, and to trigger SQL queries. However, there are now more modern and flexible options to address the workloads currently running on PythonShell. This topic explains how to migrate your workloads from AWS Glue Python shell jobs to one of these alternative options in order to take advantage of the newer capabilities that are available.

This topic explains how to migrate from AWS Glue Python shell jobs to alternative options.

Migrating workload to AWS Glue Spark jobs

AWS Glue Spark and PySpark jobs allow you to run your workloads in a distributed fashion. Since both AWS Glue Python Shell jobs and AWS Glue Spark jobs run on the same platform, it's easy to migrate, and you can continue using existing AWS Glue features that you're using with Python Shell jobs, such as AWS Glue Workflows, AWS Glue Triggers, AWS Glue's Amazon EventBridge integration, \ PIP-based package installation, and so on.

However, AWS Glue Spark jobs are designed to run Spark workloads, and the minimum number of workers is 2. If you migrate from Python Shell jobs without modifying your scripts, only one worker will be actually used, and the other workers will remain idle. This will increase your costs.

To make it efficient, rewrite your Python job script to utilize Spark's capabilities and distribute the workload across multiple workers. If your Python script is Pandas-based, it's easy to migrate using the New Pandas API on Spark. Learn more about this in the AWS Big Data Blog: Dive deep into AWS Glue 4.0 for Apache Spark.

Migrating workload to AWS Lambda

AWS Lambda is a serverless computing service that lets you run code without provisioning or managing servers. Because AWS Lambda has lower startup times and more flexible options for compute capacity, you can benefit from these advantages. For managing extra Python libraries, AWS Glue Python Shell jobs use PIP-based installation. However, for AWS Lambda, you need to choose one of the following options: a zip archive, a container image, or Lambda Layers.

On the other hand, AWS Lambda's maximum timeout is 900 seconds (15 minutes). If the job duration of your existing AWS Glue Python Shell job workload is more than that, or if your workload has a spiky pattern that may cause longer job durations, then we recommend exploring other options instead of AWS Lambda.

Migrating workload to Amazon ECS/Fargate

Amazon Elastic Container Service (Amazon ECS) is a fully managed service that simplifies the deployment, management, and scaling of containerized applications. AWS Fargate is a serverless compute engine for containerized workloads running on Amazon ECS and Amazon Elastic Kubernetes Service (Amazon EKS). There's no maximum timeout on Amazon ECS and Fargate, so this is a good option for long-running jobs. Since you have full control over your container image, you can bring your Python script and extra Python libraries into the container and use them. However, you need to containerize your Python script to use this approach.

Migrating workload to Amazon Managed Workflows for Apache Airflow Python Operator

Amazon Managed Workflows for Apache Airflow (Managed Workflows for Apache Airflow) is a managed orchestration service for Apache Airflow that makes it easier to set up and operate end-to-end data pipelines in the cloud at scale. If you already have an MWAA environment, it will be straightforward to use the Python operator instead of AWS Glue Python Shell jobs. The Python operator is an operator that runs Python code inside an Airflow workflow. However, if you don't have an existing MWAA environment, we recommend exploring other options.

Migrating workload to Amazon SageMaker AI AI training jobs

Amazon SageMaker AI Training is a fully managed machine learning (ML) service offered by Amazon SageMaker AI that helps you efficiently train a wide range of ML models at scale. The core of Amazon SageMaker AI AI jobs is the containerization of ML workloads and the capability of managing AWS compute resources. If you prefer a serverless environment where there is no maximum timeout, Amazon SageMaker AI AI training jobs could be a good fit for you. However, the startup latency tends to be longer than AWS Glue Python Shell jobs. For jobs that are latency-sensitive, we recommend exploring other options.

Monitoring AWS Glue

Monitoring is an important part of maintaining the reliability, availability, and performance of AWS Glue and your other AWS solutions. AWS provides monitoring tools that you can use to watch AWS Glue, report when something is wrong, and take action automatically when appropriate:

You can use the following automated monitoring tools to watch AWS Glue and report when something is wrong:

- **Amazon CloudWatch Events** delivers a near real-time stream of system events that describe changes in AWS resources. CloudWatch Events enables automated event-driven computing. You can write rules that watch for certain events and trigger automated actions in other AWS services when these events occur. For more information, see the Amazon CloudWatch Events User Guide.

- **Amazon CloudWatch Logs** enables you to monitor, store, and access your log files from Amazon EC2 instances, AWS CloudTrail, and other sources. CloudWatch Logs can monitor information in the log files and notify you when certain thresholds are met. You can also archive your log data in highly durable storage. For more information, see the Amazon CloudWatch Logs User Guide.
- **AWS CloudTrail** captures API calls and related events made by or on behalf of your AWS account and delivers the log files to an Amazon S3 bucket that you specify. You can identify which users and accounts call AWS, the source IP address from which the calls are made, and when the calls occur. For more information, see the AWS CloudTrail User Guide.

Additionally, you have access to the following insights in the AWS Glue console to help you debug and profile jobs:

- Spark jobs – you can see a visualization of selected CloudWatch metrics series, and newer jobs have access to the Spark UI. For more information, see the section called “Monitoring Spark jobs”.
- Ray jobs – you can see a visualization of selected CloudWatch metrics series. For more information, see the section called “Ray job metrics”.

Topics

- AWS tags in AWS Glue
- Automating AWS Glue with EventBridge
- Monitoring AWS Glue resources
- Logging AWS Glue API calls with AWS CloudTrail

AWS tags in AWS Glue

To help you manage your AWS Glue resources, you can optionally assign your own tags to some AWS Glue resource types. A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value, both of which you define. You can use tags in AWS Glue to organize and identify your resources. Tags can be used to create cost accounting reports and restrict access to resources. If you're using AWS Identity and Access Management, you can control which users in your AWS account have permission to create, edit, or delete tags. In addition to the permissions to call the tag-related APIs, you also need the `glue:GetConnection` permission to call tagging APIs on connections, and the `glue:GetDatabase` permission to call tagging APIs on databases. For more information, see ABAC with AWS Glue.

In AWS Glue, you can tag the following resources:

- Connection
- Database
- Crawler
- Interactive session
- Development endpoint
- Job
- Trigger
- Workflow
- Blueprint
- Machine learning transform
- Data quality ruleset
- Stream schemas
- Stream schema registries

> **Note**
>
> As a best practice, to allow tagging of these AWS Glue resources, always include the `glue:TagResource` action in your policies.

Consider the following when using tags with AWS Glue.

- A maximum of 50 tags are supported per entity.
- In AWS Glue, you specify tags as a list of key-value pairs in the format `{"string": "string" ...}`
- When you create a tag on an object, the tag key is required, and the tag value is optional.
- The tag key and tag value are case sensitive.
- The tag key and the tag value must not contain the prefix *aws*. No operations are allowed on such tags.
- The maximum tag key length is 128 Unicode characters in UTF-8. The tag key must not be empty or null.

- The maximum tag value length is 256 Unicode characters in UTF-8. The tag value may be empty or null.

Tagging support for AWS Glue connections

You can restrict `CreateConnection`, `UpdateConnection`, `GetConnection` and, `DeleteConnection` action permission based on the resource tag. This enables you to implement the least privilege access control on AWS Glue jobs with JDBC data sources which need to fetch JDBC connection information from the Data Catalog.

Example usage

Create an AWS Glue connection with the tag ["connection-category", "dev-test"].

Specify the tag condition for the `GetConnection` action in the IAM policy.

```
{
    "Effect": "Allow",
    "Action": [
      "glue:GetConnection"
      ],
    "Resource": "*",
    "Condition": {
      "ForAnyValue:StringEquals": {
        "aws:ResourceTag/tagKey": "dev-test"
      }
    }
 }
```

Examples

The following examples create a job with assigned tags.

AWS CLI

```
aws glue create-job --name job-test-tags --role MyJobRole --command
 Name=glueetl,ScriptLocation=S3://aws-glue-scripts//prod-job1
--tags key1=value1,key2=value2
```

AWS CloudFormation JSON

```
{
  "Description": "AWS Glue Job Test Tags",
  "Resources": {
    "MyJobRole": {
      "Type": "AWS::IAM::Role",
      "Properties": {
        "AssumeRolePolicyDocument": {
          "Version": "2012-10-17",
          "Statement": [
            {
              "Effect": "Allow",
              "Principal": {
                "Service": [
                  "glue.amazonaws.com"
                ]
              },
              "Action": [
                "sts:AssumeRole"
              ]
            }
          ]
        },
        "Path": "/",
        "Policies": [
          {
            "PolicyName": "root",
            "PolicyDocument": {
              "Version": "2012-10-17",
              "Statement": [
                {
                  "Effect": "Allow",
                  "Action": "*",
                  "Resource": "*"
                }
              ]
            }
          }
        ]
      }
    },
    "MyJob": {
      "Type": "AWS::Glue::Job",
      "Properties": {
        "Command": {
          "Name": "glueetl",
          "ScriptLocation": "s3://aws-glue-scripts//prod-job1"
        },
        "DefaultArguments": {
          "--job-bookmark-option": "job-bookmark-enable"
        },
        "ExecutionProperty": {
          "MaxConcurrentRuns": 2
        },
        "MaxRetries": 0,
        "Name": "cf-job1",
        "Role": {
           "Ref": "MyJobRole",
           "Tags": {
                "key1": "value1",
                "key2": "value2"
           }
        }
      }
    }
  }
}
```

AWS CloudFormation YAML

```
Description: AWS Glue Job Test Tags
Resources:
  MyJobRole:
    Type: AWS::IAM::Role
    Properties:
      AssumeRolePolicyDocument:
        Version: '2012-10-17'
        Statement:
          - Effect: Allow
            Principal:
              Service:
                - glue.amazonaws.com
            Action:
              - sts:AssumeRole
      Path: "/"
      Policies:
        - PolicyName: root
          PolicyDocument:
            Version: '2012-10-17'
```

```
              Statement:
                - Effect: Allow
                  Action: "*"
                  Resource: "*"
  MyJob:
    Type: AWS::Glue::Job
    Properties:
      Command:
        Name: glueetl
        ScriptLocation: s3://aws-glue-scripts//prod-job1
      DefaultArguments:
        "--job-bookmark-option": job-bookmark-enable
      ExecutionProperty:
        MaxConcurrentRuns: 2
      MaxRetries: 0
      Name: cf-job1
      Role:
        Ref: MyJobRole
        Tags:
           key1: value1
           key2: value2
```

For more information, see AWS Tagging Strategies.

For information about how to control access using tags, see ABAC with AWS Glue.

Automating AWS Glue with EventBridge

You can use Amazon EventBridge to automate your AWS services and respond automatically to system events such as application availability issues or resource changes. Events from AWS services are delivered to EventBridge in near real time. You can write simple rules to indicate which events are of interest to you, and what automated actions to take when an event matches a rule. The actions that can be automatically triggered include the following:

- Invoking an AWS Lambda function
- Invoking Amazon EC2 Run Command
- Relaying the event to Amazon Kinesis Data Streams
- Activating an AWS Step Functions state machine
- Notifying an Amazon SNS topic or an Amazon SQS queue

Some examples of using EventBridge with AWS Glue include the following:

- Activating a Lambda function when an ETL job succeeds
- Notifying an Amazon SNS topic when an ETL job fails

The following EventBridge are generated by AWS Glue.

- Events for `"detail-type":"Glue Job State Change"` are generated for `SUCCEEDED`, `FAILED`, `TIMEOUT`, and `STOPPED`.
- Events for `"detail-type":"Glue Job Run Status"` are generated for `RUNNING`, `STARTING`, and `STOPPING` job runs when they exceed the job delay notification threshold. You must set the job delay notification threshold property to receive these events.

  Only one event is generated per job run status when the job delay notification threshold is exceeded.
- Events for `"detail-type":"Glue Crawler State Change"` are generated for `Started`, `Succeeded`, and `Failed`.
- Events for `“detail_type”:“Glue Scheduled Crawler Invocation Failure”` are generated when the scheduled crawler fails to start. In the details of the notification:
  - `customerId` contains the account ID of the customer.
  - `crawlerName` contains the name of the crawler that failed to start.
  - `errorMessage` contains the exception message of the invocation failure.
- Events for `“detail_type”:“Glue Auto Statistics Invocation Failure“` are generated when the auto-managed column statistics task run fails to start. In the details of the notification:
  - `catalogId` contains the ID associated with a catalog.
  - `databaseName` contains the name of the affected database.
  - `tableName` contains the name of the affected table.
  - `errorMessage` contains the exception message of the invocation failure.
- Events for `“detail_type”:“Glue Scheduled Statistics Invocation Failure”` are generated when the (cron) scheduled column statistics task run fails to start. In the details of the notification:
  - `catalogId` contains the ID associated with a catalog.
  - `databaseName` contains the name of the affected database.

  - `tableName` contains the name of the affected table.
  - `errorMessage` contains the exception message of the invocation failure.
- Events for “`detail_type`”:“`Glue Statistics Task Started`” are generated when the column statistics task run starts.
- Events for “`detail_type`”:“`Glue Statistics Task Succeeded`” are generated when the column statistics task run succeeds.
- Events for “`detail_type`”:“`Glue Statistics Task Failed`” are generated when the column statistics task run fails.
- Events for `"detail-type":"Glue Data Catalog Database State Change"` are generated for `CreateDatabase`, `DeleteDatabase`, `CreateTable`, `DeleteTable` and `BatchDeleteTable`. For example, if a table is created or deleted, a notification is sent to EventBridge. Note that you cannot write a program that depends on the order or existence of notification events, as they might be out of sequence or missing. Events are emitted on a best effort basis. In the details of the notification:
  - The `typeOfChange` contains the name of the API operation.
  - The `databaseName` contains the name of the affected database.
  - The `changedTables` contains up to 100 names of affected tables per notification. When table names are long, multiple notifications might be created.
- Events for `"detail-type":"Glue Data Catalog Table State Change"` are generated for `UpdateTable`, `CreatePartition`, `BatchCreatePartition`, `UpdatePartition`, `DeletePartition`, `BatchUpdatePartition` and `BatchDeletePartition`. For example, if a table or partition is updated, a notification is sent to EventBridge. Note that you cannot write a program that depends on the order or existence of notification events, as they might be out of sequence or missing. Events are emitted on a best effort basis. In the details of the notification:
  - The `typeOfChange` contains the name of the API operation.
  - The `databaseName` contains the name of the database that contains the affected resources.
  - The `tableName` contains the name of the affected table.
  - The `changedPartitions` specifies up to 100 affected partitions in one notification. When partition names are long, multiple notifications might be created.

    For example if there are two partition keys, `Year` and `Month`, then `"2018,01", "2018,02"` modifies the partition where `"Year=2018" and "Month=01"` and the partition where `"Year=2018" and "Month=02"`.

```
{
    "version":"0",
    "id":"abcdef00-1234-5678-9abc-def012345678",
    "detail-type":"Glue Data Catalog Table State Change",
    "source":"aws.glue",
    "account":"123456789012",
    "time":"2017-09-07T18:57:21Z",
    "region":"us-west-2",
    "resources":["arn:aws:glue:us-west-2:123456789012:database/default/foo"],
    "detail":{
        "changedPartitions": [
            "2018,01",
            "2018,02"
        ],
        "databaseName": "default",
        "tableName": "foo",
        "typeOfChange": "BatchCreatePartition"
        }
}
```

For more information, see the Amazon CloudWatch Events User Guide. For events specific to AWS Glue, see AWS Glue Events.

Monitoring AWS Glue resources

AWS Glue has service limits to protect customers from unexpected excessive provisioning and from malicious actions intended to increase your bill. The limits also protect the service. Logging into the AWS Service Quota console, customers can view their current resource limits and request an increase (where appropriate).

AWS Glue allows you to view the service's resource usage as a percentage in Amazon CloudWatch and to configure CloudWatch alarms on it to monitor usage. Amazon CloudWatch provides monitoring for AWS resources and the customer applications running on the Amazon infrastructure. The metrics are free of charge to you. The following metrics are supported:

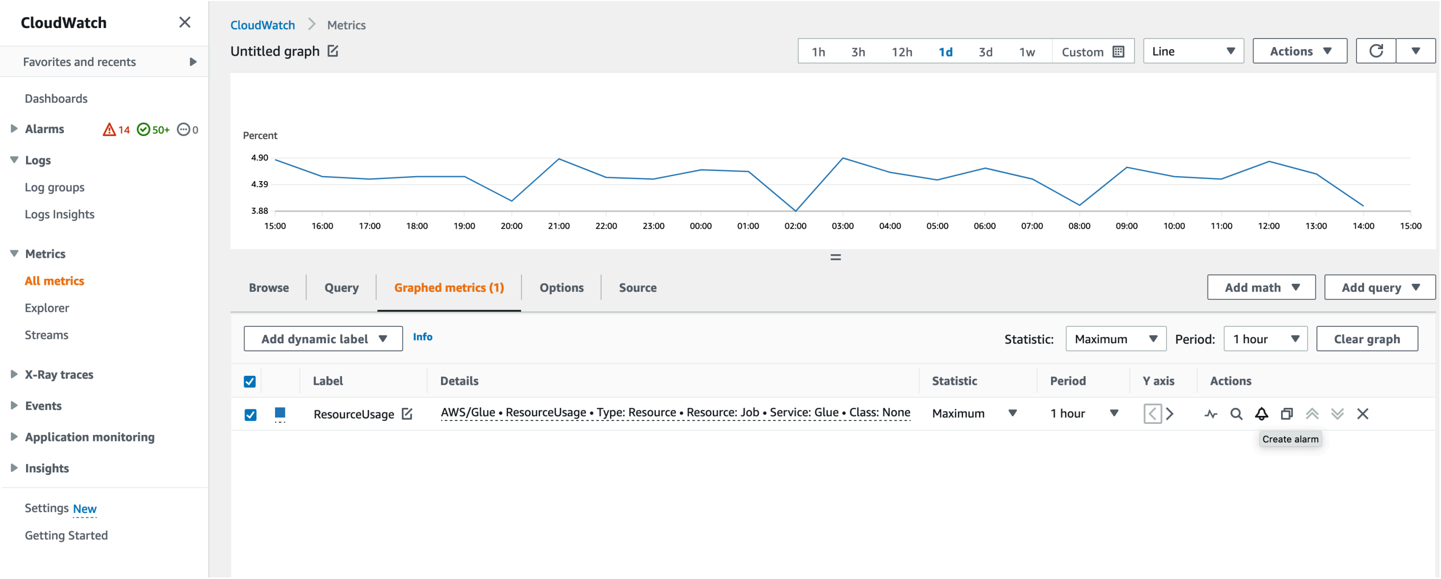
- Number of workflows per account
- Number of triggers per account
- Number of jobs per account
- Number of concurrent job runs per account

- Number of blueprints per account
- Number of interactive sessions per account

Configuring and using resource metrics

To use this feature, you can go to the Amazon CloudWatch console to view the metrics and configure alarms. The metrics are under the AWS/Glue namespace and are a percentage of the actual resource usage count divided by the resource quota. The CloudWatch metrics are delivered to your accounts, which will be no cost for you. For example, if you have 10 workflows created, and your service quota allows you to have 200 workflows in maximum, then your usage is 10/200 = 5%, and in graph, you will see a datapoint of 5 as a percentage. To be more specific:

```
Namespace: AWS/Glue
Metric name: ResourceUsage
Type: Resource
Resource: Workflow (or Trigger, Job, JobRun, Blueprint, InteractiveSession)
Service: Glue
Class: None
```


CloudWatch
Favorites and recents
Dashboards
Alarms 14 50+ 0
Logs
Log groups
Logs Insights
Metrics
All metrics
Explorer
Streams
X-Ray traces
Events
Application monitoring
Insights
Settings New
Getting Started
CloudWatch > Metrics
Untitled graph
1h 3h 12h 1d 3d 1w Custom
Line
Actions
Percent
4.90
4.39
3.88
15:00 16:00 17:00 18:00 19:00 20:00 21:00 22:00 23:00 00:00 01:00 02:00 03:00 04:00 05:00 06:00 07:00 08:00 09:00 10:00 11:00 12:00 13:00 14:00 15:00
Browse
Query
Graphed metrics (1)
Options
Source
Add math
Add query
Add dynamic label
Info
Clear graph
Label
Details
Statistic
Period
Y axis
Actions
Create alarm

Statistic: Maximum Period: 1 hour

Label	Details	Statistic	Period
ResourceUsage	AWS/Glue • ResourceUsage • Type: Resource • Resource: Job • Service: Glue • Class: None	Maximum	1 hour

To create an alarm on a metric in the CloudWatch console:

1. Once you locate the metric, go to **Graphed metrics**.
2. Click **Create alarm** under **Actions**.
3. Configure the alarm as needed.

We emit metrics whenever your resource usage has a change (such as an increase or decrease). But if your resource usage doesn't change, we emit metrics hourly, so that you have a continuous CloudWatch graph. To avoid having missing data points, we do not recommend you to configure a period less than 1 hour.

You can also configure alarms using AWS CloudFormation as in the following example. In this example, once the workflow resource usage reaches 80%, it triggers an alarm to send a message to the existing SNS topic, where you can subscribe to it to get notifications.

```
{
 "Type": "AWS::CloudWatch::Alarm",
 "Properties": {
  "AlarmName": "WorkflowUsageAlarm",
  "ActionsEnabled": true,
  "OKActions": [],
  "AlarmActions": [
   "arn:aws:sns:af-south-1:085425700061:Default_CloudWatch_Alarms_Topic"
  ],
  "InsufficientDataActions": [],
  "MetricName": "ResourceUsage",
  "Namespace": "AWS/Glue",
  "Statistic": "Maximum",
  "Dimensions": [{
    "Name": "Type",
    "Value": "Resource"
   },
   {
    "Name": "Resource",
    "Value": "Workflow"
   },
   {
    "Name": "Service",
    "Value": "Glue"
   },
   {
    "Name": "Class",
    "Value": "None"
   }
  ],
  "Period": 3600,
  "EvaluationPeriods": 1,
  "DatapointsToAlarm": 1,
  "Threshold": 80,
   "ComparisonOperator": "GreaterThanThreshold",
   "TreatMissingData": "notBreaching"
  }
 }
```

Logging AWS Glue API calls with AWS CloudTrail

AWS Glue is integrated with AWS CloudTrail, a service that provides a record of actions taken by a user, role, or an AWS service in AWS Glue. CloudTrail captures all API calls for AWS Glue as events. The calls captured include calls from the AWS Glue console and code calls to the AWS Glue API operations. If you create a trail, you can enable continuous delivery of CloudTrail events to an Amazon S3 bucket, including events for AWS Glue. If you don't configure a trail, you can still view the most recent events in the CloudTrail console in **Event history**. Using the information collected by CloudTrail, you can determine the request that was made to AWS Glue, the IP address from which the request was made, who made the request, when it was made, and additional details.

To learn more about CloudTrail, see the AWS CloudTrail User Guide.

AWS Glue information in CloudTrail

CloudTrail is enabled on your AWS account when you create the account. When activity occurs in AWS Glue, that activity is recorded in a CloudTrail event along with other AWS service events in **Event history**. You can view, search, and download recent events in your AWS account. For more information, see Viewing Events with CloudTrail Event History.

For an ongoing record of events in your AWS account, including events for AWS Glue, create a trail. A *trail* enables CloudTrail to deliver log files to an Amazon S3 bucket. By default, when you create a trail in the console, the trail applies to all AWS Regions. The trail logs events from all Regions in the AWS partition and delivers the log files to the Amazon S3 bucket that you specify. Additionally, you can configure other AWS services to further analyze and act upon the event data collected in CloudTrail logs. For more information, see the following:

- Creating a trail for your AWS account
- CloudTrail supported services and integrations
- Configuring Amazon SNS Notifications for CloudTrail
- Receiving CloudTrail Log Files from Multiple Regions and Receiving CloudTrail Log Files from Multiple Accounts

All AWS Glue actions are logged by CloudTrail and are documented in the AWS Glue API . For example, calls to the `CreateDatabase`, `CreateTable` and `CreateScript` actions generate entries in the CloudTrail log files.

Every event or log entry contains information about who generated the request. The identity information helps you determine the following:

- Whether the request was made with root or IAM user credentials.
- Whether the request was made with temporary security credentials for a role or federated user.
- Whether the request was made by another AWS service.

For more information, see the CloudTrail userIdentity Element.

However, CloudTrail doesn't log all information regarding calls. For example, it doesn't log certain sensitive information, such as the `ConnectionProperties` used in connection requests, and it logs a `null` instead of the responses returned by the following APIs:

```
BatchGetPartition       GetCrawlers          GetJobs          GetTable
CreateScript            GetCrawlerMetrics    GetJobRun        GetTables
GetCatalogImportStatus  GetDatabase          GetJobRuns       GetTableVersions
GetClassifier           GetDatabases         GetMapping       GetTrigger
GetClassifiers          GetDataflowGraph     GetObjects       GetTriggers
GetConnection           GetDevEndpoint       GetPartition     GetUserDefinedFunction
GetConnections          GetDevEndpoints      GetPartitions    GetUserDefinedFunctions
GetCrawler              GetJob               GetPlan
```

Understanding AWS Glue log file entries

A trail is a configuration that enables delivery of events as log files to an Amazon S3 bucket that you specify. CloudTrail log files contain one or more log entries. An event represents a single request from any source and includes information about the requested action, the date and time of the action, request parameters, and so on. CloudTrail log files aren't an ordered stack trace of the public API calls, so they don't appear in any specific order.

The following example shows a CloudTrail log entry that demonstrates the `DeleteCrawler` action.

```
{
  "eventVersion": "1.05",
  "userIdentity": {
    "type": "IAMUser",
    "principalId": "AKIAIOSFODNN7EXAMPLE",
    "arn": "arn:aws:iam::123456789012:user/johndoe",
    "accountId": "123456789012",
    "accessKeyId": "AKIAIOSFODNN7EXAMPLE",
    "userName": "johndoe"
  },
  "eventTime": "2017-10-11T22:29:49Z",
  "eventSource": "glue.amazonaws.com",
  "eventName": "DeleteCrawler",
  "awsRegion": "us-east-1",
  "sourceIPAddress": "72.21.198.64",
  "userAgent": "aws-cli/1.11.148 Python/3.6.1 Darwin/16.7.0 botocore/1.7.6",
  "requestParameters": {
    "name": "tes-alpha"
  },
  "responseElements": null,
  "requestID": "b16f4050-aed3-11e7-b0b3-75564a46954f",
  "eventID": "e73dd117-cfd1-47d1-9e2f-d1271cad838c",
  "eventType": "AwsApiCall",
  "recipientAccountId": "123456789012"
}
```

This example shows a CloudTrail log entry that demonstrates a `CreateConnection` action.

```
{
  "eventVersion": "1.05",
  "userIdentity": {
    "type": "IAMUser",
    "principalId": "AKIAIOSFODNN7EXAMPLE",
    "arn": "arn:aws:iam::123456789012:user/johndoe",
    "accountId": "123456789012",
    "accessKeyId": "AKIAIOSFODNN7EXAMPLE",
    "userName": "johndoe"
  },
  "eventTime": "2017-10-13T00:19:19Z",
  "eventSource": "glue.amazonaws.com",
  "eventName": "CreateConnection",
  "awsRegion": "us-east-1",
  "sourceIPAddress": "72.21.198.66",
  "userAgent": "aws-cli/1.11.148 Python/3.6.1 Darwin/16.7.0 botocore/1.7.6",
  "requestParameters": {
    "connectionInput": {
      "name": "test-connection-alpha",
      "connectionType": "JDBC",
      "physicalConnectionRequirements": {
        "subnetId": "subnet-323232",
        "availabilityZone": "us-east-1a",
        "securityGroupIdList": [
          "sg-12121212"
        ]
      }
    }
  },
  "responseElements": null,
  "requestID": "27136ebc-afac-11e7-a7d6-ab217e5c3f19",
  "eventID": "e8b3baeb-c511-4597-880f-c16210c60a4a",
  "eventType": "AwsApiCall",
  "recipientAccountId": "123456789012"
}
```

AWS Glue Streaming

AWS Glue Streaming, a component of AWS Glue, enables you to efficiently handle streaming data in near real-time, empowering you to carry out crucial tasks such as data ingestion, processing, and machine learning. Using the Apache Spark Streaming framework, AWS Glue Streaming provides a serverless service that can handle streaming data at scale. AWS Glue provides various optimizations on top of Apache Spark such as serverless infrastructure, auto-scaling, visual job development, instant-on notebooks for streaming jobs and other performance improvements.

Use cases for streaming

Some common use cases for AWS Glue Streaming include:

Near-real-time data processing: AWS Glue Streaming allows organizations to process streaming data in near real-time, enabling them to derive insights and make timely decisions based on the latest information.

Fraud detection: You can utilize AWS Glue Streaming for real-time analysis of streaming data, making it valuable for detecting fraudulent activities, such as credit card fraud, network intrusion, or online scams. By continuously processing and analyzing incoming data, you can swiftly identify suspicious patterns or anomalies.

Social media analytics: AWS Glue Streaming can process real-time social media data, such as tweets, posts, or comments, enabling organizations to monitor trends, sentiment analysis, and manage brand reputation in real-time.

Internet of Things (IoT) analytics: AWS Glue Streaming is suitable for handling and analyzing high-velocity streams of data generated by IoT devices, sensors, and connected machinery. It allows for real-time monitoring, anomaly detection, predictive maintenance, and other IoT analytics use cases.

Clickstream analysis: AWS Glue Streaming can process and analyze real-time clickstream data from websites or mobile applications. This enables businesses to gain insights into user behavior, personalize user experiences, and optimize marketing campaigns based on real-time clickstream data.

Log monitoring and analysis: AWS Glue Streaming can continuously process and analyze log data from servers, applications, or network devices in real-time. This helps in detecting anomalies, troubleshooting issues, and monitoring system health and performance.

Recommendation systems: AWS Glue Streaming can process user activity data in real-time and update recommendation models dynamically. This allows for personalized and real-time recommendations based on user behavior and preferences.

These are some examples of the diverse range of use cases where AWS Glue Streaming can be applied. Its integration with the AWS ecosystem and managed services make it a convenient choice for real-time stream processing and analytics in the cloud.

What are the benefits of using AWS Glue Streaming?

The benefits of using AWS Glue Streaming are as follows:

- **Serverless**: AWS Glue Streaming is serverless, eliminating the need to manage infrastructure. This reduces the operational overhead and allows users to focus on data processing and analytics tasks rather than infrastructure management.
- **Autoscaling**: AWS Glue Streaming provides autoscaling capabilities, dynamically adjusting the processing capacity based on the workload. It automatically scales out or in to handle fluctuations in data volume, ensuring optimal performance and resource utilization.
- **Visual development**: Streaming job development can be complex. AWS Glue Streaming addresses this challenge by offering AWS Glue Studio, a visual authoring tool. AWS Glue Studio simplifies the process of creating streaming workflows and enables developers to design and manage streaming applications visually, reducing the learning curve and increasing productivity.
- **Cost-effective**: As a serverless service, AWS Glue Streaming offers cost efficiency by eliminating the need for provisioning and maintaining infrastructure. Users are billed based on the resources consumed during the execution of streaming jobs, allowing for cost optimization and scaling based on actual usage.
- **Handles complex workloads**: AWS Glue Streaming is designed to handle complex streaming workloads. It can process and analyze large volumes of real-time data, support advanced transformations, and integrate with other AWS services, enabling sophisticated streaming data pipelines and analytics workflows.
- **No lock-in**: AWS Glue Streaming provides flexibility and avoids vendor lock-in. Users can leverage AWS Glue Streaming as part of the broader AWS ecosystem, integrating it with other AWS services seamlessly. This allows for easy integration with existing data sources, applications, and services without being tied to a specific technology or platform.

When to use AWS Glue Streaming?

There are many options when it comes to streaming use cases. We recommend AWS Glue streaming in the following scenarios.

1. **If you are already using AWS Glue or Spark for batch processing**, AWS Glue Streaming is the ideal choice for you. It provides a seamless transition to building streaming jobs without the need to learn a new language or framework. Leveraging your existing knowledge and infrastructure, AWS Glue Streaming simplifies the job development process and allows you to easily extend your data processing capabilities to real-time streaming scenarios.
2. **If you require a unified service or product to handle batch, streaming, and event-driven workloads**, AWS Glue Streaming is the solution for you. With AWS Glue Streaming, you can consolidate your data processing needs into a single framework, eliminating the complexity of managing multiple systems. This enables efficient development and maintenance of diverse data workflows while ensuring consistency and compatibility across different workload types.
3. **AWS Glue Streaming is well-suited for scenarios involving extremely large streaming data volumes and complex transformations**, such as joins between streams or relational databases. It can efficiently process and analyze massive streams of data, enabling you to tackle demanding workloads with ease. Whether it is high-velocity data ingestion or intricate data manipulations, AWS Glue Streaming's scalability and advanced processing capabilities ensure optimal performance and accurate results.
4. **If you prefer a visual approach to building streaming jobs**, AWS Glue offers AWS Glue Studio, with which you can visually design and manage your streaming applications, simplifying the development process. This intuitive interface enables developers to create, configure, and monitor streaming workflows using a visual interface, reducing the learning curve and increasing productivity.
5. **AWS Glue Streaming is an excellent choice for near-real-time use cases where there are stringent SLAs (Service Level Agreements)** greater than 10 seconds.
6. **If you are building a transactional data lake using Apache Iceberg, Apache Hudi, or Delta Lake**, AWS Glue Streaming provides native support for these open table formats. This seamless integration enables you to process streaming data directly from these transactional data lakes, ensuring data consistency, integrity, and compatibility.
7. **When needing to ingest streaming data for a variety of data targets**: AWS Glue Streaming provides native targets to a variety of data targets such as Amazon Redshift, Amazon RDS, Amazon Aurora, Oracle, SQL Server and other targets.

Supported data sources

AWS Glue Streaming supports the following data sources:

- Amazon Kinesis
- Amazon MSK (Managed Streaming for Apache Kafka)
- Self-managed Apache Kafka

Supported data targets

AWS Glue Streaming supports a variety of data targets such as:

- Data targets supported by AWS Glue Data Catalog
- Amazon S3
- Amazon Redshift
- MySQL
- PostgreSQL
- Oracle
- Microsoft SQL Server
- Snowflake
- Any database that can be connected using JDBC
- Apache Iceberg, Delta and Apache Hudi
- AWS Glue Marketplace connectors

Tutorial: Build your first streaming workload using AWS Glue Studio

In this tutorial, you are going to learn how to create a streaming job using AWS Glue Studio. AWS Glue Studio is a visual interface to create AWS Glue jobs.

You can create streaming extract, transform, and load (ETL) jobs that run continuously and consume data from streaming sources in Amazon Kinesis Data Streams, Apache Kafka, and Amazon Managed Streaming for Apache Kafka (Amazon MSK).

Prerequisites

To follow this tutorial you'll need a user with AWS console permissions to use AWS Glue, Amazon Kinesis, Amazon S3, Amazon Athena, AWS CloudFormation, AWS Lambda and Amazon Cognito.

Consume streaming data from Amazon Kinesis

Topics

- Generating mock data with Kinesis Data Generator
- Creating an AWS Glue streaming job with AWS Glue Studio
- Performing a transformation and storing the transformed result in Amazon S3

Generating mock data with Kinesis Data Generator

You can synthetically generate sample data in JSON format using the Kinesis Data Generator (KDG). You can find full instructions and details in the tool documentation.

1. To get started, click

   Launch Stack

   to run an AWS CloudFormation template on your AWS environment.

   > **Note**
   >
   > You may encounter a CloudFormation template failure because some resources, such as the Amazon Cognito user for Kinesis Data Generator already exist in your AWS account. This could be because you already set that up from another tutorial or blog. To address this, you can either try the template in a new AWS account for a fresh start, or explore a different AWS Region. These options let you run the tutorial without conflicting with existing resources.

   The template provisions a Kinesis data stream and a Kinesis Data Generator account for you. It also creates an Amazon S3 bucket to hold the data and a Glue Service Role with the required permission for this tutorial.

2. Enter a **Username** and **Password** that the KDG will use to authenticate. Note the username and password for further usage.

3. Select **Next** all the way to the last step. Acknowledge the creation of IAM resources. Check for any errors at the top of the screen, such as the password not meeting the minimum requirements, and deploy the template.
4. Navigate to the **Outputs** tab of the stack. Once the template is deployed, it will display the generated property **KinesisDataGeneratorUrl**. Click that URL.
5. Enter the **Username** and **Password** you noted down.
6. Select the Region you are using and select the Kinesis Stream `GlueStreamTest-{AWS::AccountId}`
7. Enter the following template:

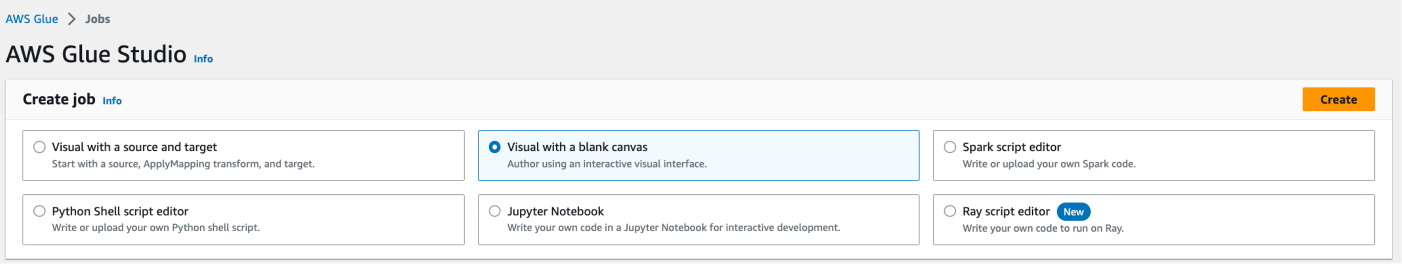
```
{
    "ventilatorid": {{random.number(100)}},
    "eventtime": "{{date.now("YYYY-MM-DD HH:mm:ss")}}",
    "serialnumber": "{{random.uuid}}",
    "pressurecontrol": {{random.number(
        {
            "min":5,
            "max":30
        }
    )}},
    "o2stats": {{random.number(
        {
            "min":92,
            "max":98
        }
    )}},
    "minutevolume": {{random.number(
        {
            "min":5,
            "max":8
        }
    )}},
    "manufacturer": "{{random.arrayElement(
        ["3M", "GE","Vyaire", "Getinge"]
    )}}"
}
```

   You can now view mock data with **Test template** and ingest the mock data to Kinesis with **Send data**.

8. Click **Send data** and generate 5-10K records to Kinesis.

Creating an AWS Glue streaming job with AWS Glue Studio

1. Navigate to AWS Glue in the console on the same Region.
2. Select **ETL jobs** under the left side navigation bar under **Data Integration and ETL**.
3. Create an AWS Glue Job via **Visual with a blank canvas**.



○ Visual with a source and target
Start with a source, ApplyMapping transform, and target.

● Visual with a blank canvas
Author using an interactive visual interface.

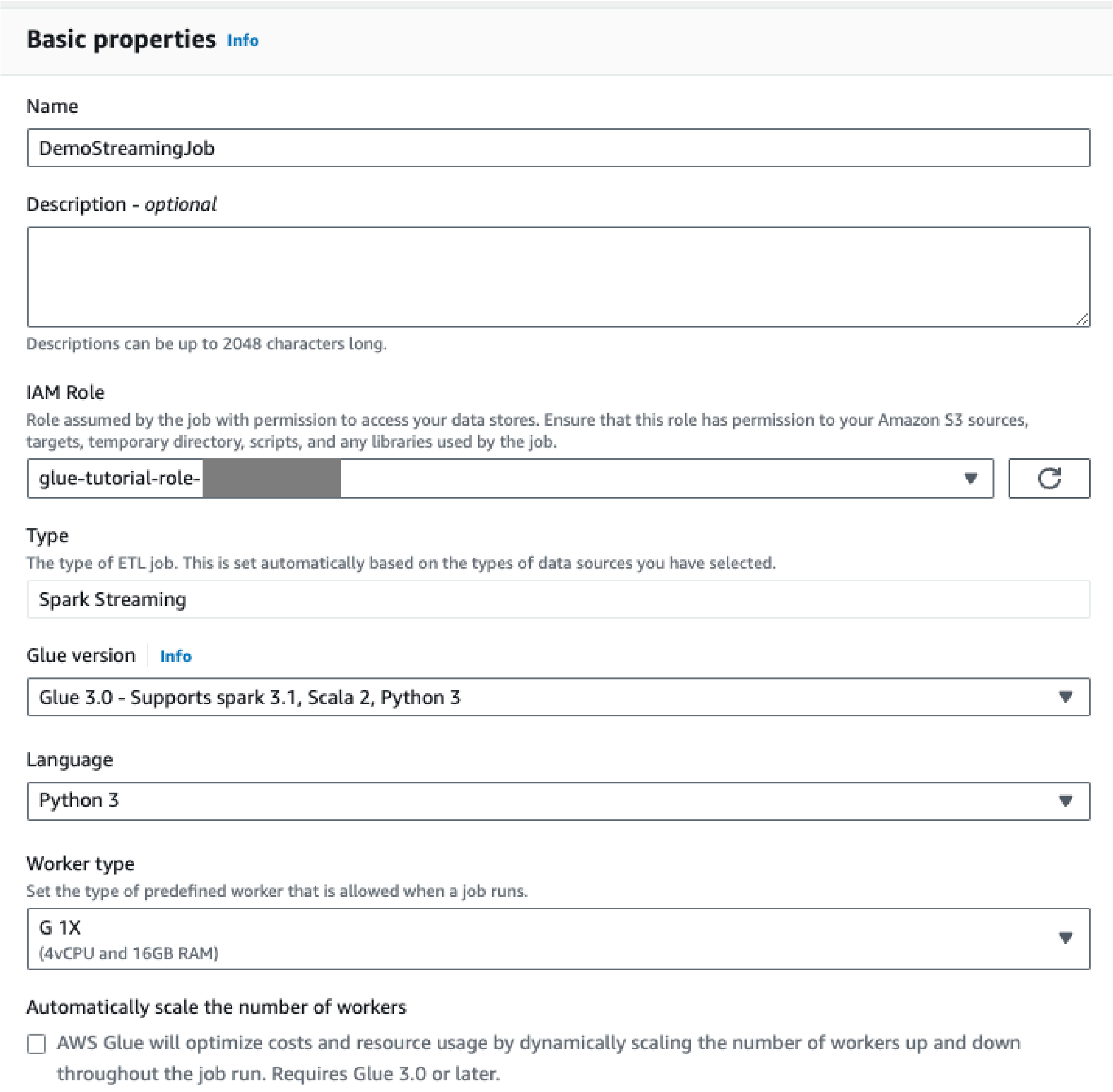
○ Spark script editor
Write or upload your own Spark code.

○ Python Shell script editor
Write or upload your own Python shell script.

○ Jupyter Notebook
Write your own code in a Jupyter Notebook for interactive development.

○ Ray script editor
Write your own code to run on Ray.

4. Navigate to the **Job Details** tab.
5. For the AWS Glue job name, enter `DemoStreamingJob`.
6. For **IAM Role**, select the role provisioned by the CloudFormation template, `glue-tutorial-role-${AWS::AccountId}`.
7. For **Glue version**, select **Glue 3.0**. Leave all other options as default.



Basic properties Info

Name

DemoStreamingJob

Description - *optional*

Descriptions can be up to 2048 characters long.

IAM Role

Role assumed by the job with permission to access your data stores. Ensure that this role has permission to your Amazon S3 sources, targets, temporary directory, scripts, and any libraries used by the job.

glue-tutorial-role-[redacted]

Type

The type of ETL job. This is set automatically based on the types of data sources you have selected.

Spark Streaming

Glue version Info

Glue 3.0 - Supports spark 3.1, Scala 2, Python 3

Language

Python 3

Worker type

Set the type of predefined worker that is allowed when a job runs.

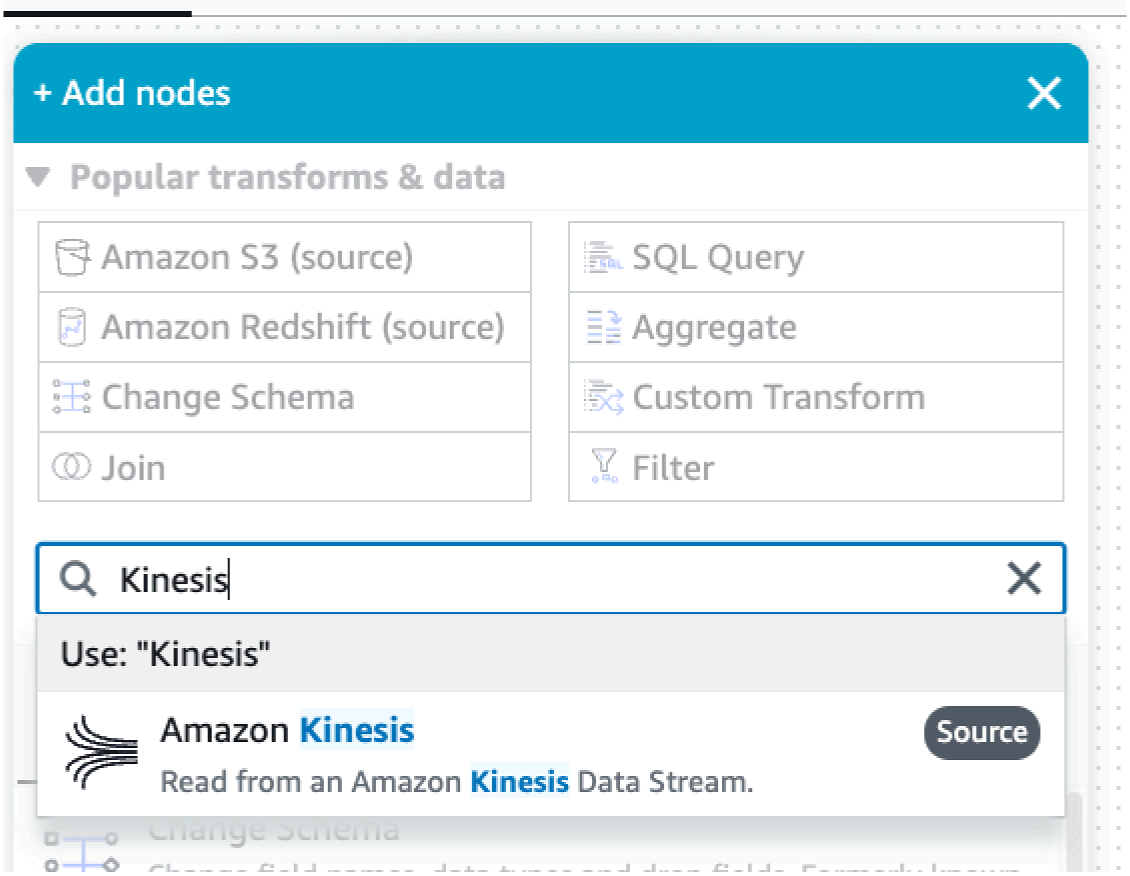
G 1X
(4vCPU and 16GB RAM)

Automatically scale the number of workers

☐ AWS Glue will optimize costs and resource usage by dynamically scaling the number of workers up and down throughout the job run. Requires Glue 3.0 or later.

8. Navigate to the **Visual tab**.

9. Click on the plus icon. Enter **Kinesis** in the search bar. Select the **Amazon Kinesis** data source.


+ Add nodes
Popular transforms & data
Amazon S3 (source)
SQL Query
Amazon Redshift (source)
Aggregate
Change Schema
Custom Transform
Join
Filter
Use: "Kinesis"
Amazon Kinesis
Source
Change Schema

Kinesis

Read from an Amazon Kinesis Data Stream.

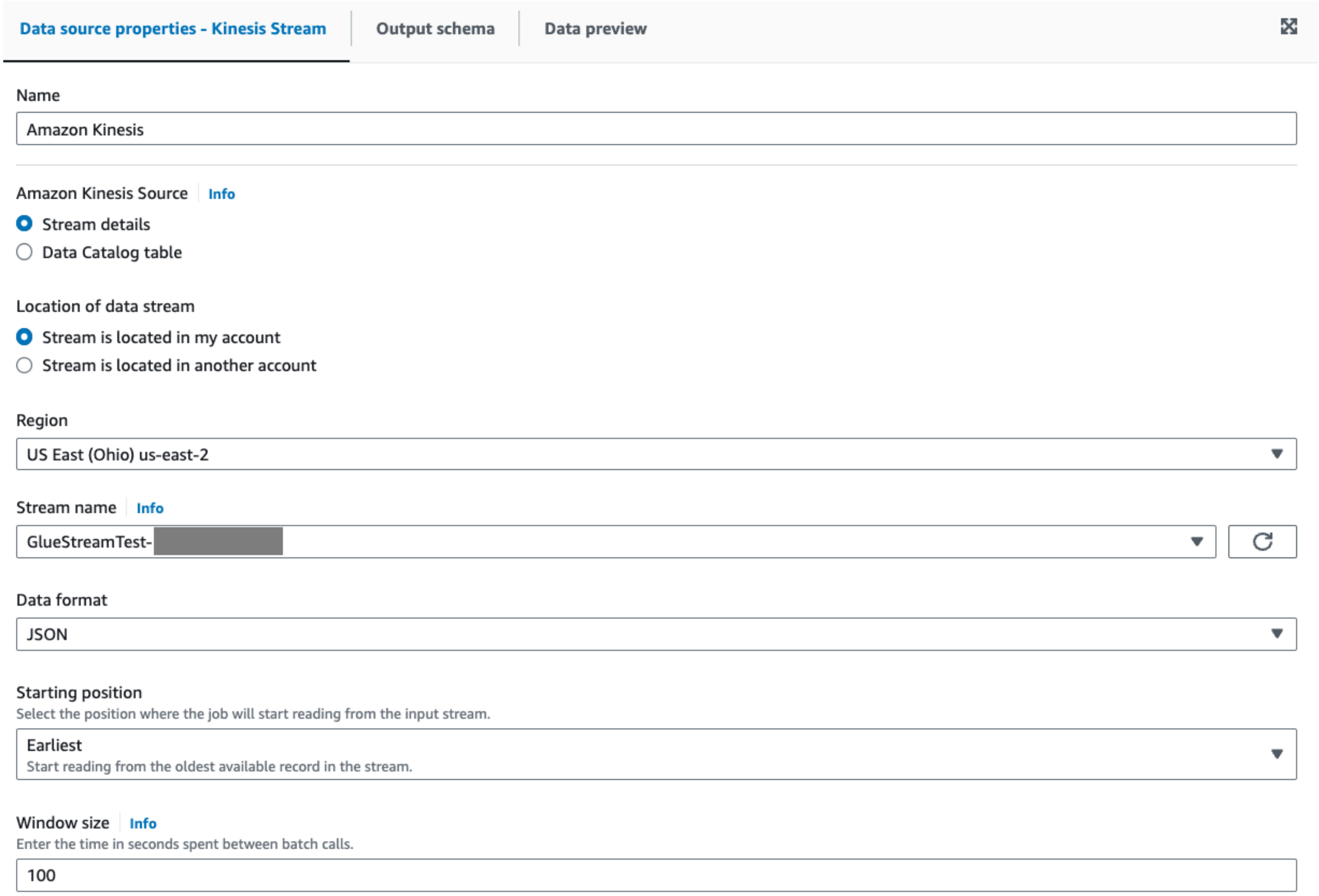
10. Select **Stream details** for **Amazon Kinesis Source** under the tab **Data source properties - Kinesis Stream**.

11. Select **Stream is located in my account** for **Location of data stream**.

12. Select the Region you are using.

13. Select the `GlueStreamTest-{AWS::AccountId}` stream.

14. Keep all other settings as default.



Name

Amazon Kinesis

Amazon Kinesis Source

● Stream details

○ Data Catalog table

Location of data stream

● Stream is located in my account

○ Stream is located in another account

Region

US East (Ohio) us-east-2

Stream name

GlueStreamTest-[redacted]

Data format

JSON

Starting position

Select the position where the job will start reading from the input stream.

Earliest

Start reading from the oldest available record in the stream.

Window size

Enter the time in seconds spent between batch calls.

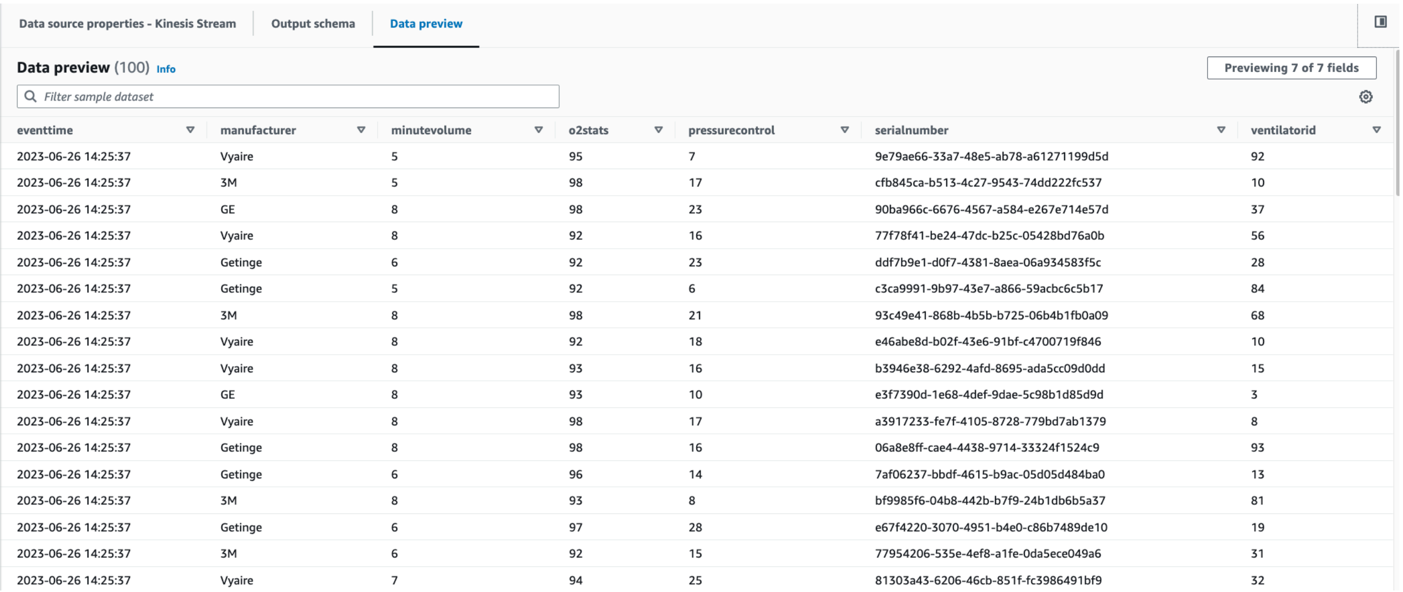
100

15. Navigate to the **Data preview** tab.

16. Click **Start data preview session**, which previews the mock data generated by KDG. Pick the Glue Service Role you previously created for the AWS Glue Streaming job.

    It takes 30-60 seconds for the preview data to show up. If it shows **No data to display**, click the gear icon and change the **Number of rows to sample** to `100`.

    You can see the sample data as below:

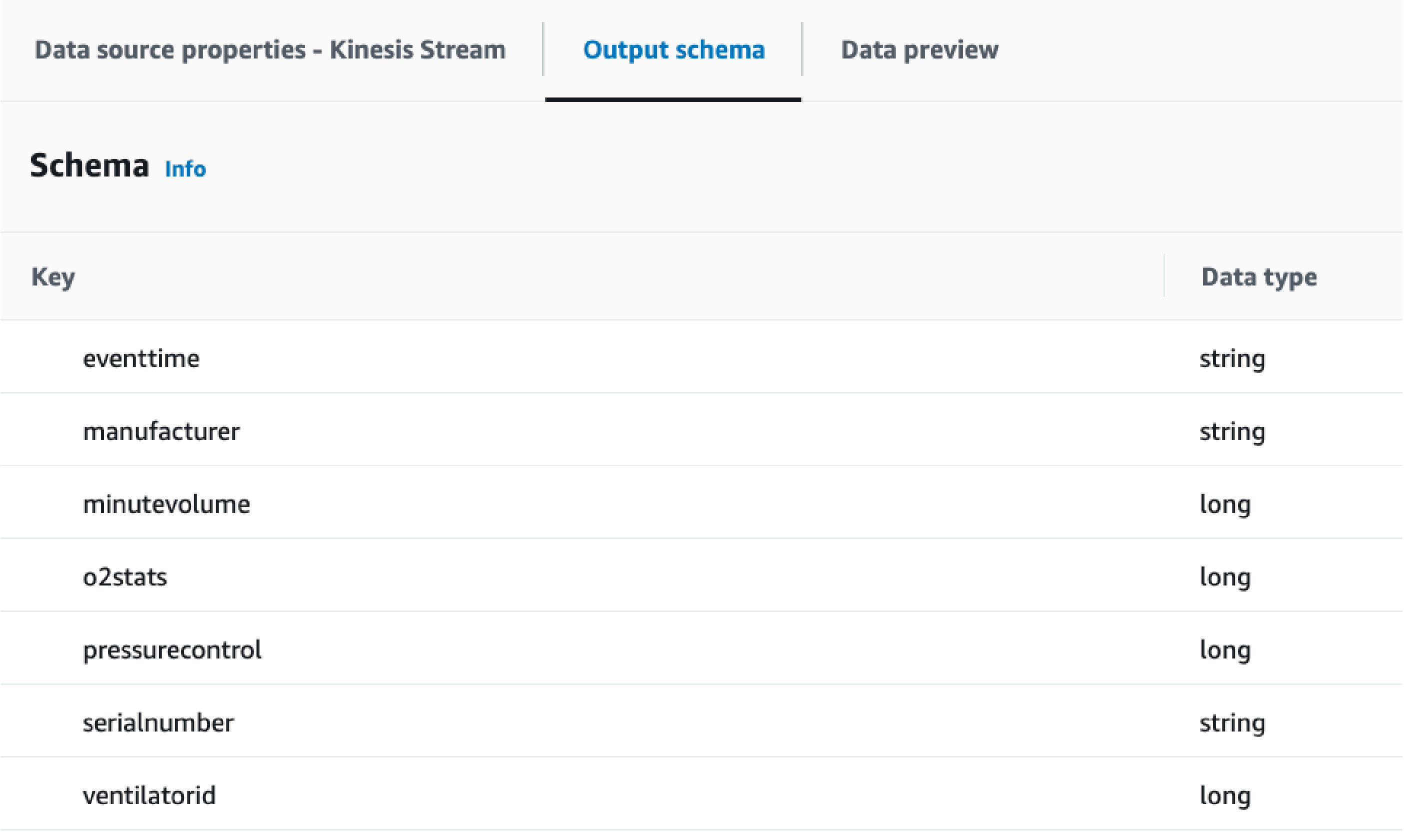


Data preview (100) Info

Filter sample dataset

eventtime	manufacturer	minutevolume	o2stats	pressurecontrol	serialnumber	ventilatorid
2023-06-26 14:25:37	Vyaire	5	95	7	9e79ae66-33a7-48e5-ab78-a61271199d5d	92
2023-06-26 14:25:37	3M	5	98	17	cfb845ca-b513-4c27-9543-74dd222fc537	10
2023-06-26 14:25:37	GE	8	98	23	90ba966c-6676-4567-a584-e267e714e57d	37
2023-06-26 14:25:37	Vyaire	8	92	16	77f78f41-be24-47dc-b25c-05428bd76a0b	56
2023-06-26 14:25:37	Getinge	6	92	23	ddf7b9e1-d0f7-4381-8aea-06a934583f5c	28
2023-06-26 14:25:37	Getinge	5	92	6	c3ca9991-9b97-43e7-a866-59acbc6c5b17	84
2023-06-26 14:25:37	3M	8	98	21	93c49e41-868b-4b5b-b725-06b4b1fb0a09	68
2023-06-26 14:25:37	Vyaire	8	92	18	e46abe8d-b02f-43e6-91bf-c4700719f846	10
2023-06-26 14:25:37	Vyaire	8	93	16	b3946e38-6292-4afd-8695-ada5cc09d0dd	15
2023-06-26 14:25:37	GE	8	93	10	e3f7390d-1e68-4def-9dae-5c98b1d85d9d	3
2023-06-26 14:25:37	Vyaire	8	98	17	a3917233-fe7f-4105-8728-779bd7ab1379	8
2023-06-26 14:25:37	Getinge	8	98	16	06a8e8ff-cae4-4438-9714-33324f1524c9	93
2023-06-26 14:25:37	Getinge	6	96	14	7af06237-bbdf-4615-b9ac-05d05d484ba0	13
2023-06-26 14:25:37	3M	8	93	8	bf9985f6-04b8-442b-b7f9-24b1db6b5a37	81
2023-06-26 14:25:37	Getinge	6	97	28	e67f4220-3070-4951-b4e0-c86b7489de10	19
2023-06-26 14:25:37	3M	6	92	15	77954206-535e-4ef8-a1fe-0da5ece049a6	31
2023-06-26 14:25:37	Vyaire	7	94	25	81303a43-6206-46cb-851f-fc3986491bf9	32

You can also see the inferred schema in the **Output schema** tab.

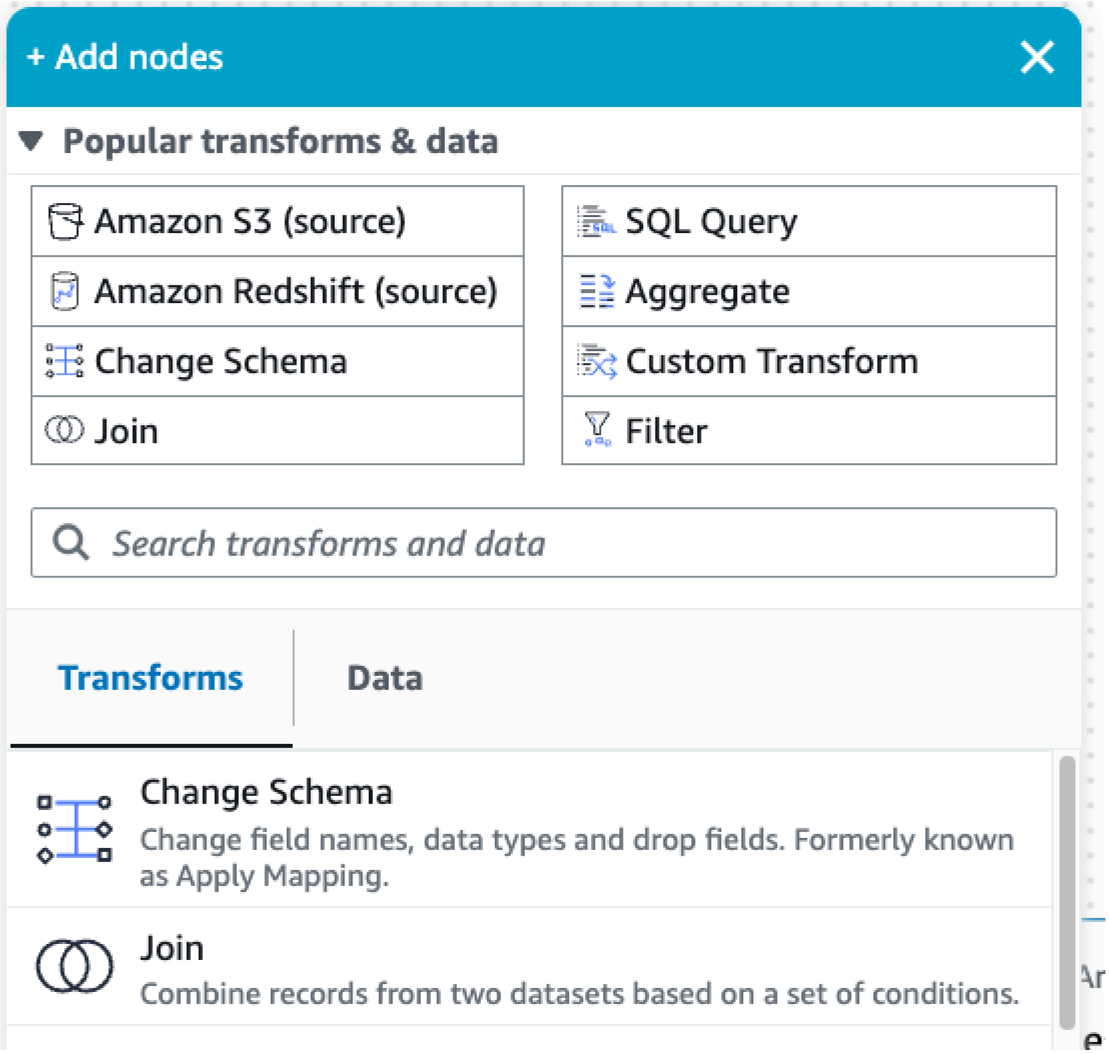


Schema Info

Key	Data type
eventtime	string
manufacturer	string
minutevolume	long
o2stats	long
pressurecontrol	long
serialnumber	string
ventilatorid	long

Performing a transformation and storing the transformed result in Amazon S3

1. With the source node selected, click on the plus icon on the top left to add a **Transforms** step.
2. Select the **Change Schema** step.



+ Add nodes
Popular transforms & data
Amazon S3 (source)
SQL Query
Amazon Redshift (source)
Aggregate
Change Schema
Custom Transform
Join
Filter
Search transforms and data
Transforms
Data

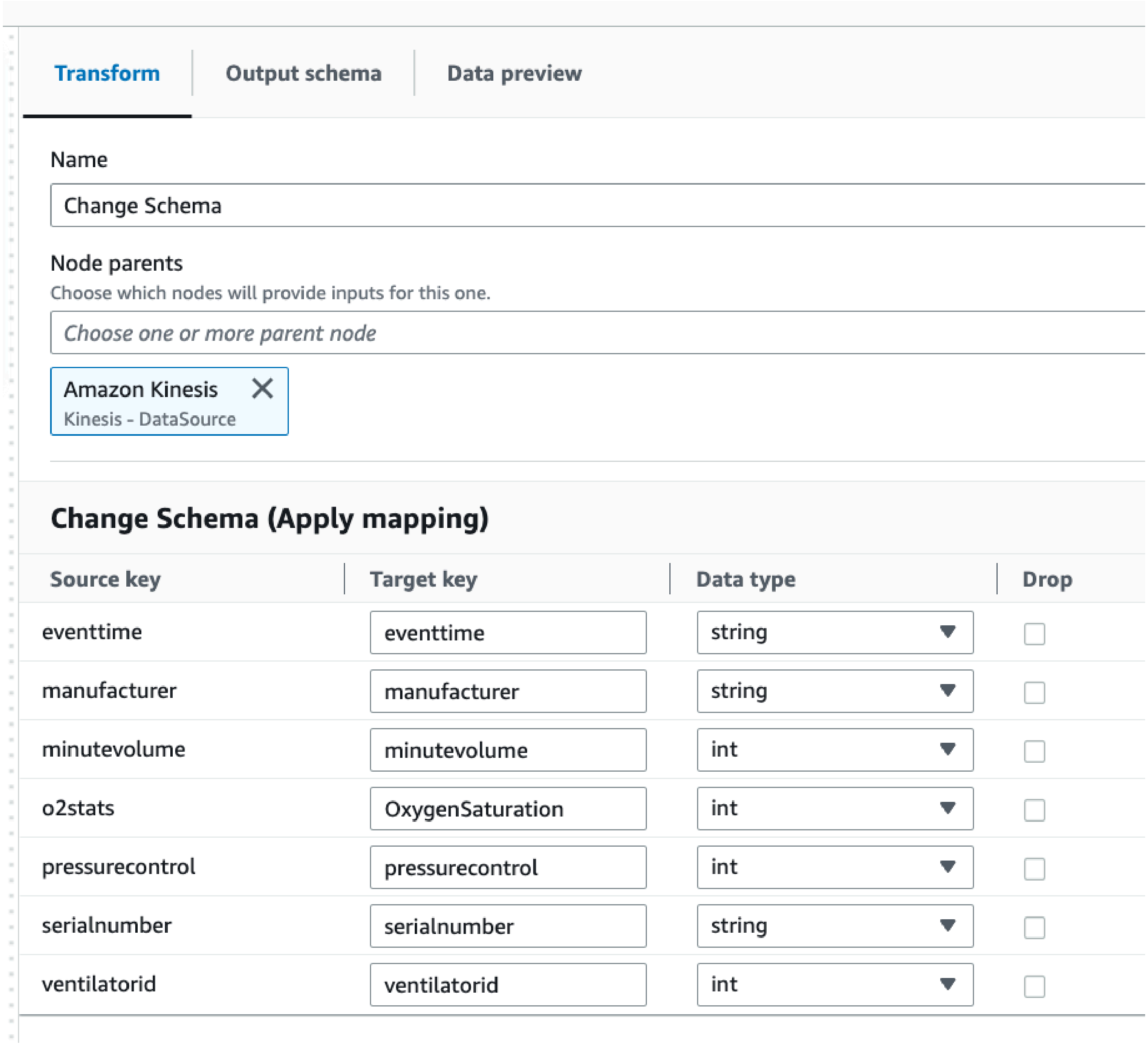
Change Schema

Change field names, data types and drop fields. Formerly known as Apply Mapping.

Join

Combine records from two datasets based on a set of conditions.

3. You can rename fields and convert the data type of fields in this step. Rename the `o2stats` column to `OxygenSaturation` and convert all `long` data type to `int`.



Name

Change Schema

Node parents

Choose which nodes will provide inputs for this one.

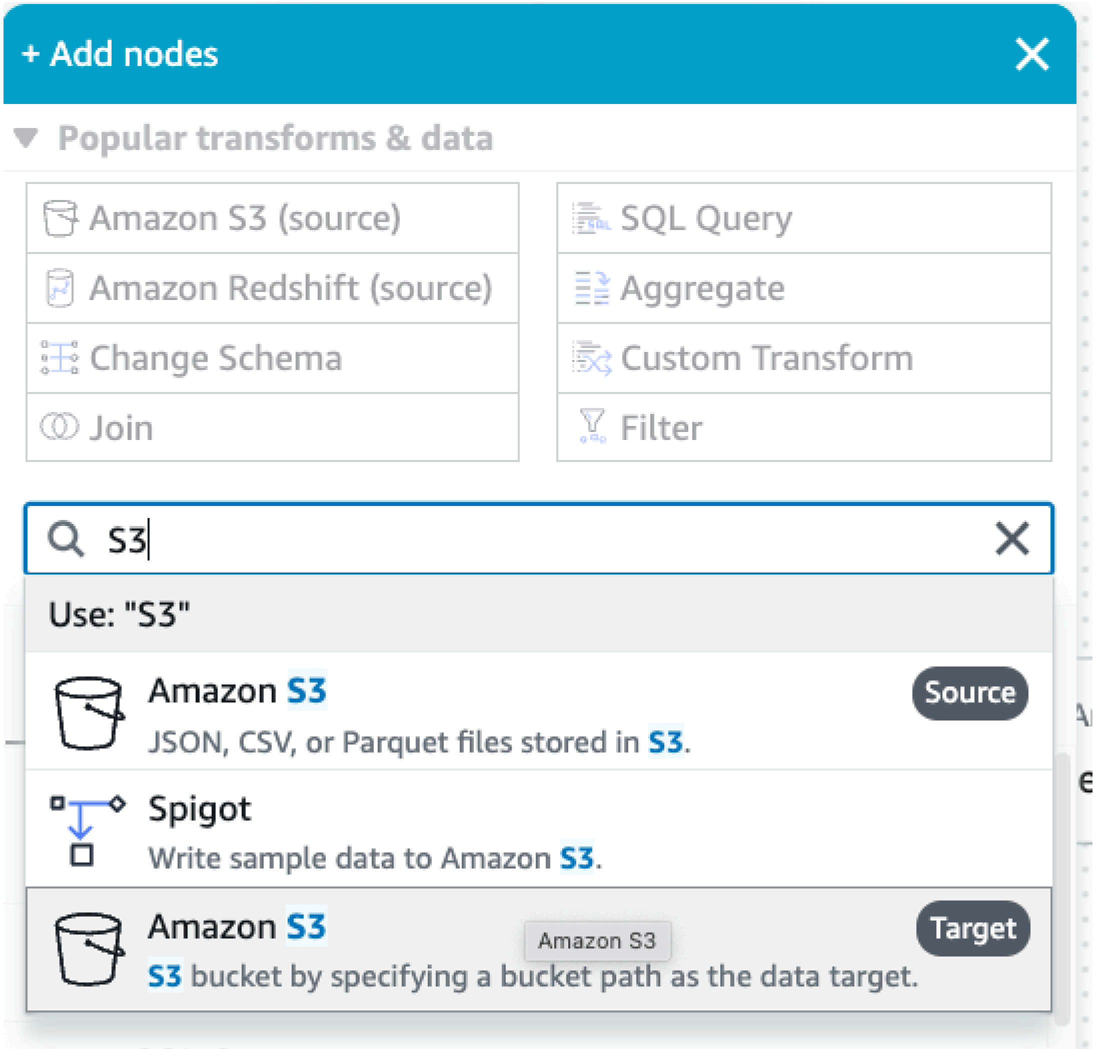
Choose one or more parent node

Amazon Kinesis
Kinesis - DataSource

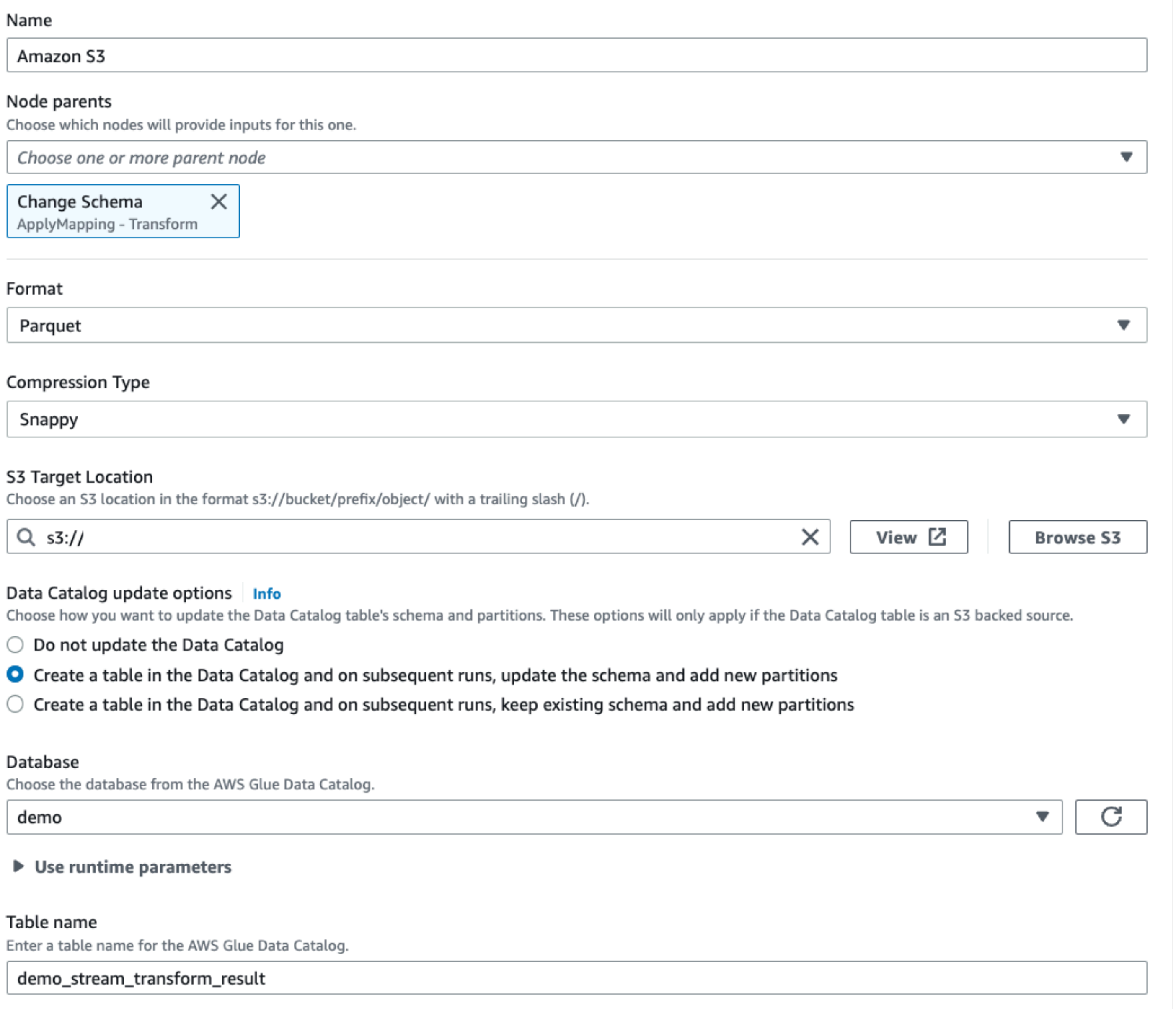
Change Schema (Apply mapping)

Source key	Target key	Data type	Drop
eventtime	eventtime	string	☐
manufacturer	manufacturer	string	☐
minutevolume	minutevolume	int	☐
o2stats	OxygenSaturation	int	☐
pressurecontrol	pressurecontrol	int	☐
serialnumber	serialnumber	string	☐
ventilatorid	ventilatorid	int	☐

4. Click on the plus icon to add an **Amazon S3** target. Enter S3 in the search box and select the **Amazon S3 - Target** transform step.


+ Add nodes
Popular transforms & data
Amazon S3 (source)
Amazon Redshift (source)
Change Schema
Join
SQL Query
Aggregate
Custom Transform
Filter
S3
Use: "S3"
Amazon S3
Source
JSON, CSV, or Parquet files stored in S3.
Spigot
Write sample data to Amazon S3.
Amazon S3
Amazon S3
Target
S3 bucket by specifying a bucket path as the data target.

5. Select **Parquet** as the target file format.
6. Select **Snappy** as the compression type.
7. Enter an **S3 Target Location** created by the CloudFormation template, `streaming-tutorial-s3-target-{AWS::AccountId}`.
8. Select to **Create a table in the Data Catalog and on subsequent runs, update the schema and add new partitions**.
9. Enter the target **Database** and **Table** name to store the schema of the Amazon S3 target table.



Name

Amazon S3

Node parents

Choose which nodes will provide inputs for this one.

Choose one or more parent node

Change Schema
ApplyMapping - Transform

Format

Parquet

Compression Type

Snappy

S3 Target Location

Choose an S3 location in the format s3://bucket/prefix/object/ with a trailing slash (/).

s3://

Data Catalog update options Info

Choose how you want to update the Data Catalog table's schema and partitions. These options will only apply if the Data Catalog table is an S3 backed source.

○ Do not update the Data Catalog

● Create a table in the Data Catalog and on subsequent runs, update the schema and add new partitions

○ Create a table in the Data Catalog and on subsequent runs, keep existing schema and add new partitions

Database

Choose the database from the AWS Glue Data Catalog.

demo

Table name

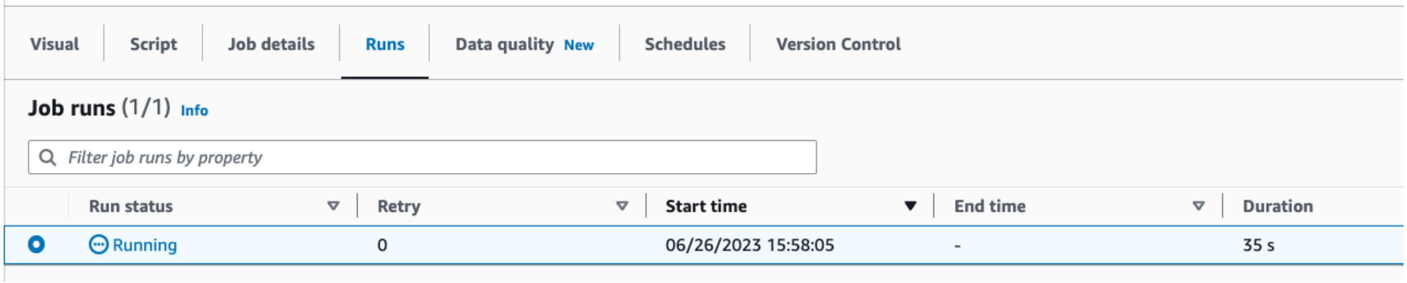
Enter a table name for the AWS Glue Data Catalog.

demo_stream_transform_result

10. Click on the **Script** tab to view the generated code.

11. Click **Save** on the top right to save the ETL code and then click **Run** to kick-off the AWS Glue streaming job.

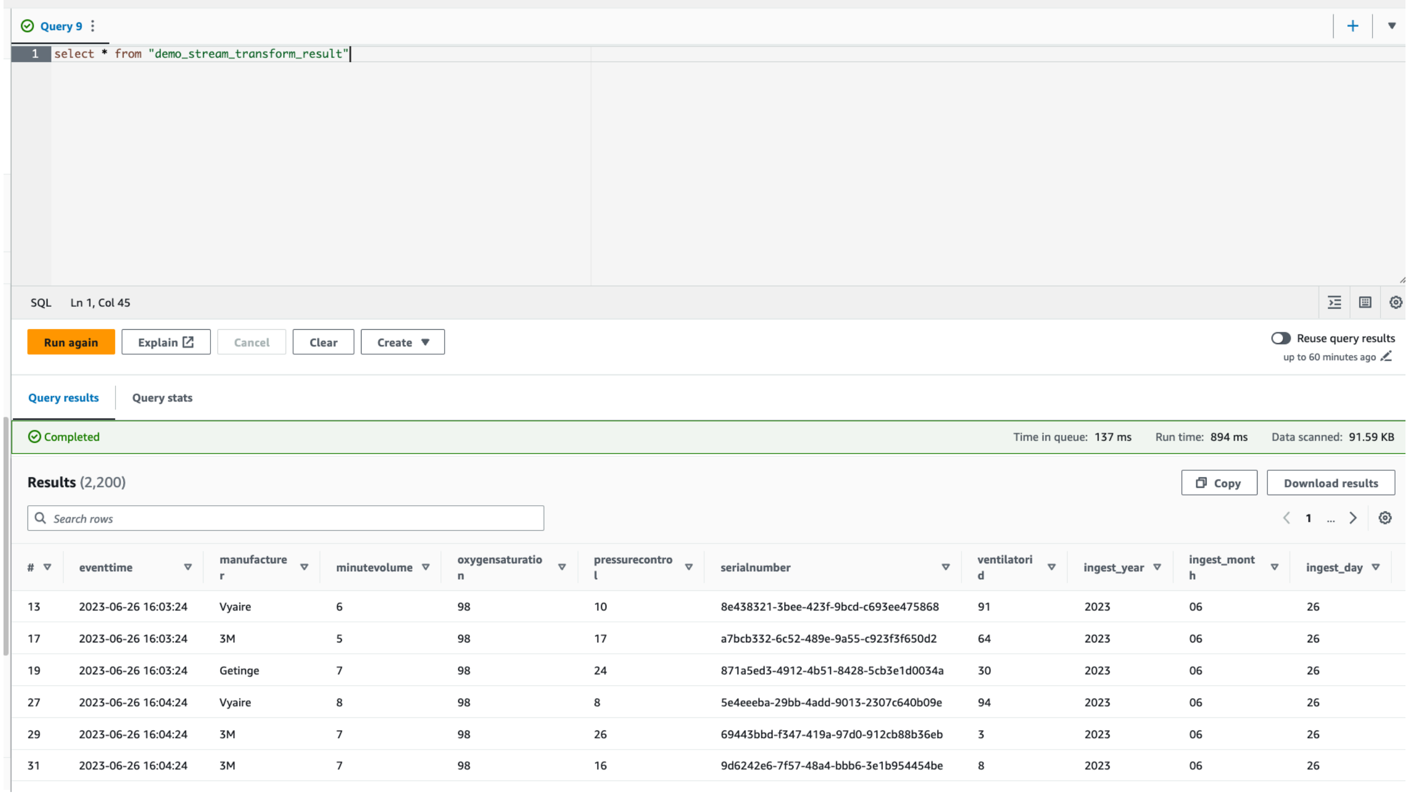
    You can find the **Run status** in the **Runs** tab. Let the job run for 3-5 minutes and then stop the job.


Visual
Script
Job details
Runs
Data quality New
Schedules
Version Control
Info
Filter job runs by property

Job runs (1/1)

	Run status	Retry	Start time	End time	Duration
●	Running	0	06/26/2023 15:58:05	-	35 s

12.Verify the new table created in Amazon Athena.


Query 9
SQL Ln 1, Col 45
Run again
Explain
Cancel
Clear
Create
Reuse query results
up to 60 minutes ago
Query results
Query stats
Copy
Download results
Search rows
1 ...

```
select * from "demo_stream_transform_result"
```

Completed Time in queue: 137 ms Run time: 894 ms Data scanned: 91.59 KB

Results (2,200)

#	eventtime	manufacturer	minutevolume	oxygensaturation	pressurecontrol	serialnumber	ventilatorid	ingest_year	ingest_month	ingest_day
13	2023-06-26 16:03:24	Vyaire	6	98	10	8e438321-3bee-423f-9bcd-c693ee475868	91	2023	06	26
17	2023-06-26 16:03:24	3M	5	98	17	a7bcb332-6c52-489e-9a55-c923f3f650d2	64	2023	06	26
19	2023-06-26 16:03:24	Getinge	7	98	24	871a5ed3-4912-4b51-8428-5cb3e1d0034a	30	2023	06	26
27	2023-06-26 16:04:24	Vyaire	8	98	8	5e4eeeba-29bb-4add-9013-2307c640b09e	94	2023	06	26
29	2023-06-26 16:04:24	3M	7	98	26	69443bbd-f347-419a-97d0-912cb88b36eb	3	2023	06	26
31	2023-06-26 16:04:24	3M	7	98	16	9d6242e6-7f57-48a4-bbb6-3e1b954454be	8	2023	06	26

Tutorial: Build your first streaming workload using AWS Glue Studio notebooks

In this tutorial, you will explore how to leverage AWS Glue Studio notebooks to interactively build and refine your ETL jobs for near real-time data processing. Whether you're new to AWS Glue or looking to enhance your skill set, this guide will walk you through the process, empowering you to harness the full potential of AWS Glue interactive session notebooks.

With AWS Glue Streaming, you can create streaming extract, transform, and load (ETL) jobs that run continuously and consume data from streaming sources such as Amazon Kinesis Data Streams, Apache Kafka, and Amazon Managed Streaming for Apache Kafka (Amazon MSK).

Prerequisites

To follow this tutorial you'll need a user with AWS console permissions to use AWS Glue, Amazon Kinesis, Amazon S3, Amazon Athena, AWS CloudFormation, AWS Lambda and Amazon Cognito.

Consume streaming data from Amazon Kinesis

Topics

- Generating mock data with Kinesis Data Generator
- Creating an AWS Glue streaming job with AWS Glue Studio
- Clean up
- Conclusion

Generating mock data with Kinesis Data Generator

Note

If you have already completed our previous Tutorial: Build your first streaming workload using AWS Glue Studio, you already have the Kinesis Data Generator installed on your account and you can skip steps 1-8 below and move on to the section Creating an AWS Glue streaming job with AWS Glue Studio.

You can synthetically generate sample data in JSON format using the Kinesis Data Generator (KDG). You can find full instructions and details in the tool documentation.

1. To get started, click

   Launch Stack

   to run an AWS CloudFormation template on your AWS environment.

> **Note**
>
> You may encounter a CloudFormation template failure because some resources, such as the Amazon Cognito user for Kinesis Data Generator already exist in your AWS account. This could be because you already set that up from another tutorial or blog. To address this, you can either try the template in a new AWS account for a fresh start, or explore a different AWS Region. These options let you run the tutorial without conflicting with existing resources.

The template provisions a Kinesis data stream and a Kinesis Data Generator account for you.

2. Enter a **Username** and **Password** that the KDG will use to authenticate. Note the username and password for further usage.
3. Select **Next** all the way to the last step. Acknowledge the creation of IAM resources. Check for any errors at the top of the screen, such as the password not meeting the minimum requirements, and deploy the template.
4. Navigate to the **Outputs** tab of the stack. Once the template is deployed, it will display the generated property **KinesisDataGeneratorUrl**. Click that URL.
5. Enter the **Username** and **Password** you noted down.
6. Select the Region you are using and select the Kinesis Stream `GlueStreamTest-{AWS::AccountId}`
7. Enter the following template:

```
{
    "ventilatorid": {{random.number(100)}},
    "eventtime": "{{date.now("YYYY-MM-DD HH:mm:ss")}}",
    "serialnumber": "{{random.uuid}}",
    "pressurecontrol": {{random.number(
        {
            "min":5,
            "max":30
        }
    )}},
    "o2stats": {{random.number(
        {
            "min":92,
            "max":98
        }
    )}},
    "minutevolume": {{random.number(
        {
            "min":5,
            "max":8
        }
    )}},
    "manufacturer": "{{random.arrayElement(
        ["3M", "GE","Vyaire", "Getinge"]
    )}}"
}
```

You can now view mock data with **Test template** and ingest the mock data to Kinesis with **Send data**.

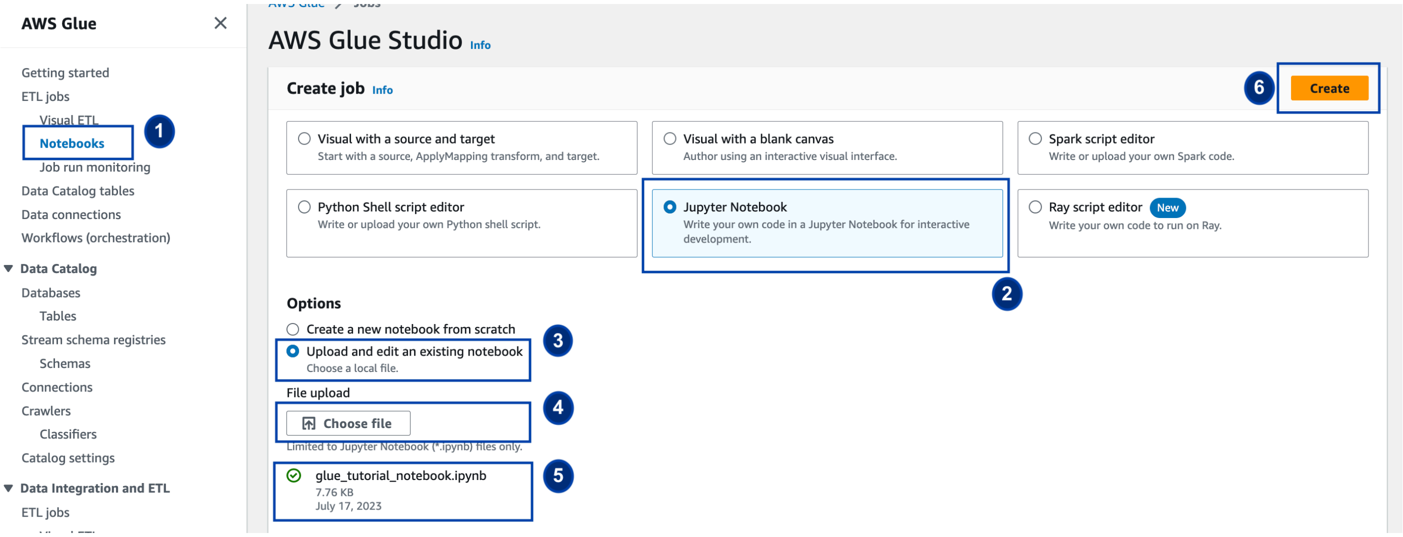
8. Click **Send data** and generate 5-10K records to Kinesis.

Creating an AWS Glue streaming job with AWS Glue Studio

AWS Glue Studio is a visual interface that simplifies the process of designing, orchestrating, and monitoring data integration pipelines. It enables users to build data transformation pipelines without writing extensive code. Apart from the visual job authoring experience, AWS Glue Studio also includes a Jupyter notebook backed by AWS Glue Interactive sessions, which you will be using in the remainder of this tutorial.

Set up the AWS Glue Streaming interactive sessions job

1. Download the provided notebook file and save it to a local directory
2. Open the AWS Glue Console and on the left pane click **Notebooks** > **Jupyter Notebook** > **Upload and edit an existing notebook**. Upload the notebook from the previous step and click **Create**.


AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
1
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
ETL jobs
AWS Glue Studio Info
Create job Info
6
Create
2
3
4
5

Create job

○ Visual with a source and target
Start with a source, ApplyMapping transform, and target.

○ Visual with a blank canvas
Author using an interactive visual interface.

○ Spark script editor
Write or upload your own Spark code.

○ Python Shell script editor
Write or upload your own Python shell script.

● Jupyter Notebook
Write your own code in a Jupyter Notebook for interactive development.

○ Ray script editor New
Write your own code to run on Ray.

Options

○ Create a new notebook from scratch

● Upload and edit an existing notebook
Choose a local file.

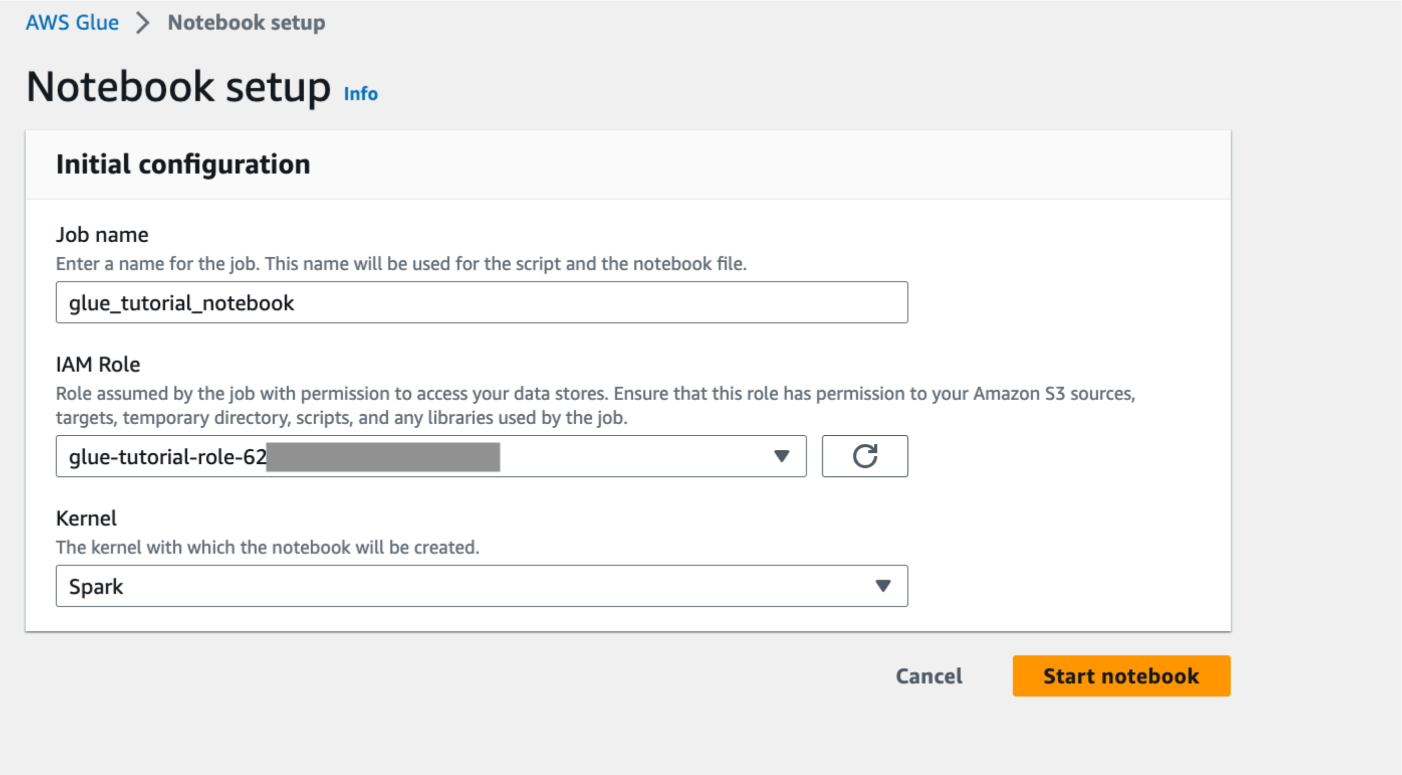
File upload

Choose file

Limited to Jupyter Notebook (*.ipynb) files only.

glue_tutorial_notebook.ipynb
7.76 KB
July 17, 2023

3. Provide the job a name, role and select the default Spark kernel. Next click **Start notebook**. For the **IAM Role**, select the role provisioned by the CloudFormation template. You can see this in the **Outputs** tab of CloudFormation.


AWS Glue > Notebook setup
Notebook setup Info
Cancel
Start notebook

Initial configuration

Job name
Enter a name for the job. This name will be used for the script and the notebook file.
glue_tutorial_notebook

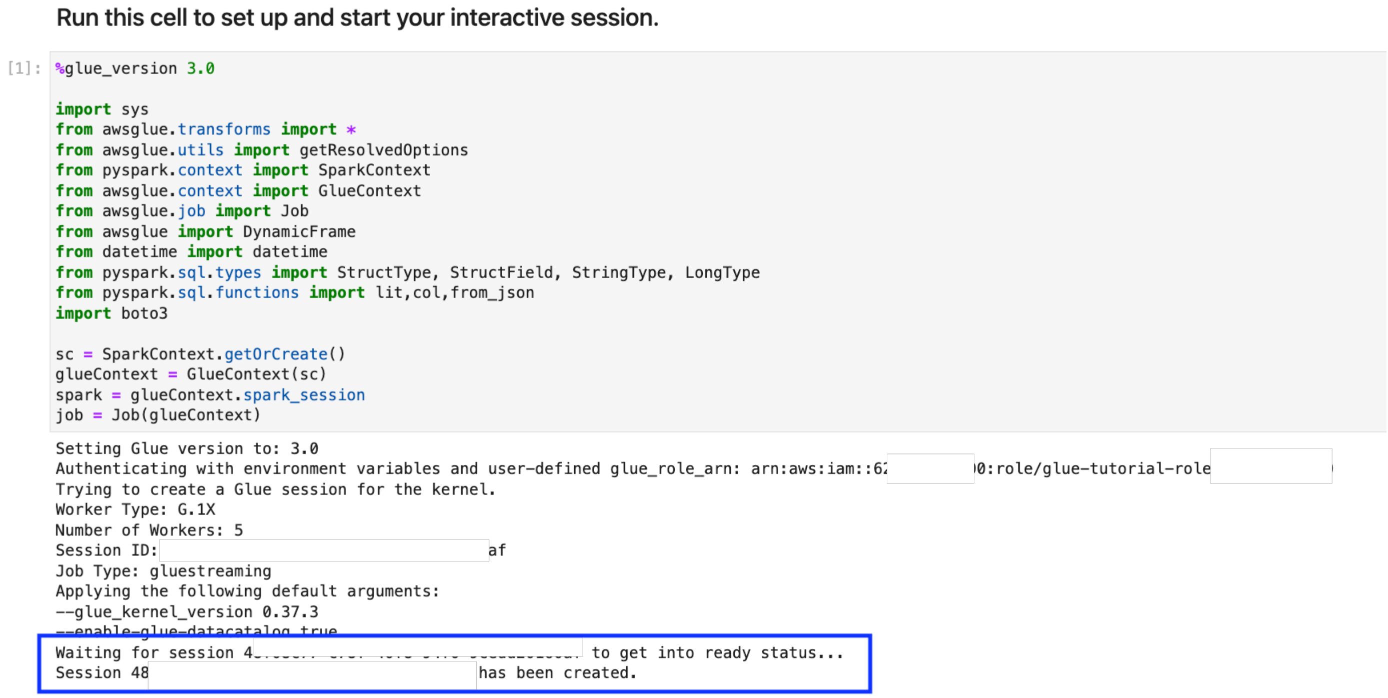
IAM Role
Role assumed by the job with permission to access your data stores. Ensure that this role has permission to your Amazon S3 sources, targets, temporary directory, scripts, and any libraries used by the job.
glue-tutorial-role-62[redacted]

Kernel
The kernel with which the notebook will be created.
Spark

The notebook has all necessary instructions to continue the tutorial. You can either run the instructions on the notebook or follow along with this tutorial to continue with the job development.

Run the notebook cells

1. (Optional) The first code cell, `%help` lists all available notebook magics. You can skip this cell for now, but feel free to explore it.
2. Start with the next code block `%streaming`. This magic sets the job type to streaming which lets you develop, debug and deploy an AWS Glue streaming ETL job.
3. Run the next cell to create an AWS Glue interactive session. The output cell has a message that confirms the session creation.

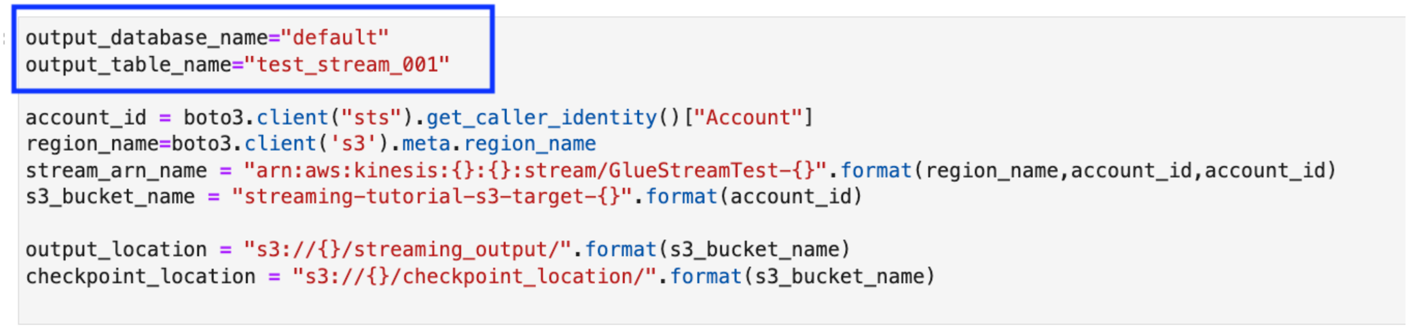


Run this cell to set up and start your interactive session.

```
[1]: %glue_version 3.0

import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job
from awsglue import DynamicFrame
from datetime import datetime
from pyspark.sql.types import StructType, StructField, StringType, LongType
from pyspark.sql.functions import lit,col,from_json
import boto3

sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
```

```
Setting Glue version to: 3.0
Authenticating with environment variables and user-defined glue_role_arn: arn:aws:iam::62[redacted]0:role/glue-tutorial-role[redacted]
Trying to create a Glue session for the kernel.
Worker Type: G.1X
Number of Workers: 5
Session ID:[redacted]af
Job Type: gluestreaming
Applying the following default arguments:
--glue_kernel_version 0.37.3
--enable-glue-datacatalog true
Waiting for session 4[redacted] to get into ready status...
Session 48[redacted]has been created.
```

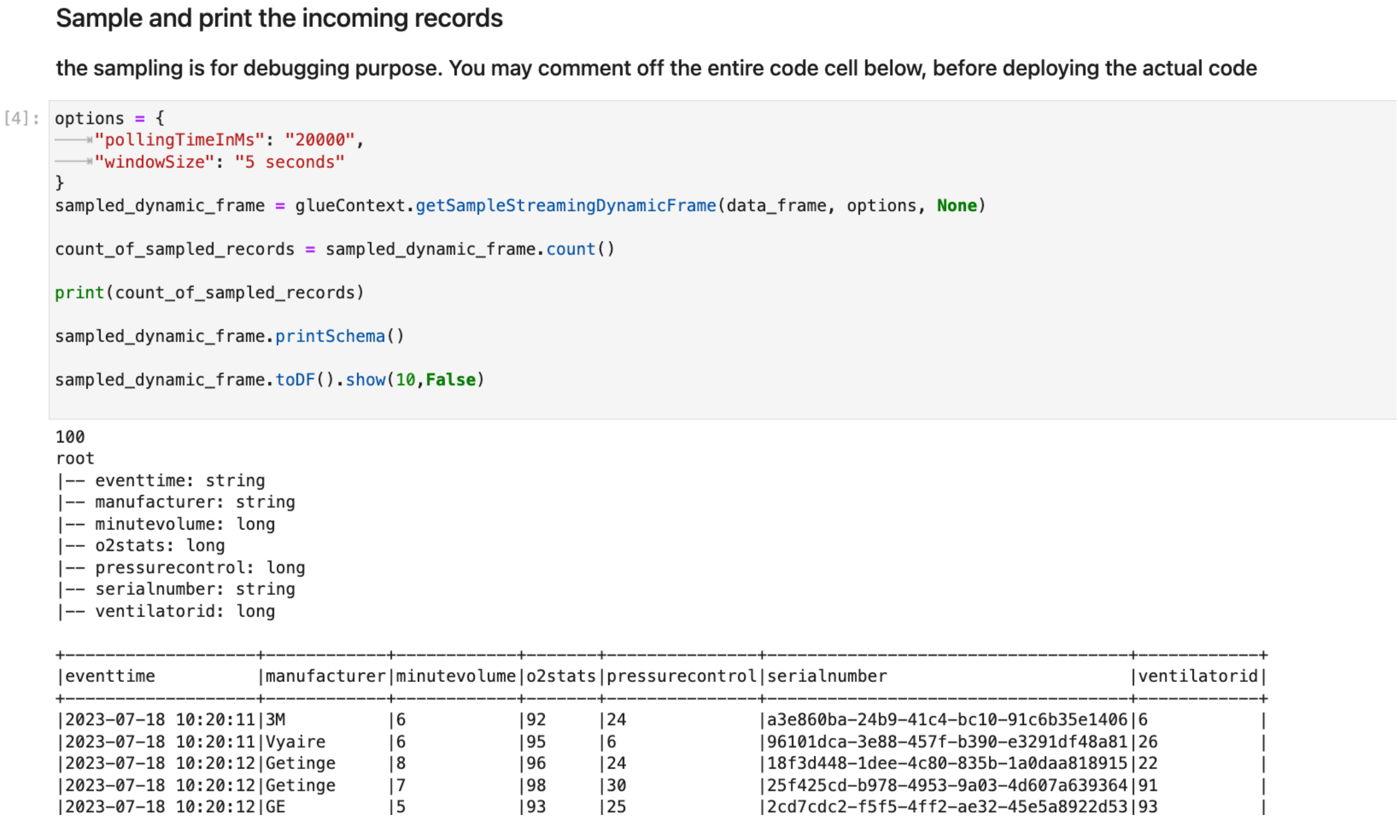
4. The next cell defines the variables. Replace the values with ones appropriate to your job and run the cell. For example:



```
output_database_name="default"
output_table_name="test_stream_001"

account_id = boto3.client("sts").get_caller_identity()["Account"]
region_name=boto3.client('s3').meta.region_name
stream_arn_name = "arn:aws:kinesis:{}:{}:stream/GlueStreamTest-{}".format(region_name,account_id,account_id)
s3_bucket_name = "streaming-tutorial-s3-target-{}".format(account_id)

output_location = "s3://{}/streaming_output/".format(s3_bucket_name)
checkpoint_location = "s3://{}/checkpoint_location/".format(s3_bucket_name)
```

5. Since the data is being streamed already to Kinesis Data Streams, your next cell will consume the results from the stream. Run the next cell. Since there are no print statements, there is no expected output from this cell.

6. In the following cell, you explore the incoming stream by taking a sample set and print its schema and the actual data. For example:



Sample and print the incoming records

the sampling is for debugging purpose. You may comment off the entire code cell below, before deploying the actual code

[4]:

```
options = {
    "pollingTimeInMs": "20000",
    "windowSize": "5 seconds"
}
sampled_dynamic_frame = glueContext.getSampleStreamingDynamicFrame(data_frame, options, None)

count_of_sampled_records = sampled_dynamic_frame.count()

print(count_of_sampled_records)

sampled_dynamic_frame.printSchema()

sampled_dynamic_frame.toDF().show(10,False)
```

```
100
root
|-- eventtime: string
|-- manufacturer: string
|-- minutevolume: long
|-- o2stats: long
|-- pressurecontrol: long
|-- serialnumber: string
|-- ventilatorid: long

+-------------------+------------+------------+-------+---------------+------------------------------------+------------+
|eventtime          |manufacturer|minutevolume|o2stats|pressurecontrol|serialnumber                        |ventilatorid|
+-------------------+------------+------------+-------+---------------+------------------------------------+------------+
|2023-07-18 10:20:11|3M          |6           |92     |24             |a3e860ba-24b9-41c4-bc10-91c6b35e1406|6           |
|2023-07-18 10:20:11|Vyaire      |6           |95     |6              |96101dca-3e88-457f-b390-e3291df48a81|26          |
|2023-07-18 10:20:12|Getinge     |8           |96     |24             |18f3d448-1dee-4c80-835b-1a0daa818915|22          |
|2023-07-18 10:20:12|Getinge     |7           |98     |30             |25f425cd-b978-4953-9a03-4d607a639364|91          |
|2023-07-18 10:20:12|GE          |5           |93     |25             |2cd7cdc2-f5f5-4ff2-ae32-45e5a8922d53|93          |
```

7. Next, define the actual data transformation logic. The cell consists of the `processBatch` method that is triggered during every micro-batch. Run the cell. At a high level, we do the following to the incoming stream:

    a. Select a subset of the input columns.

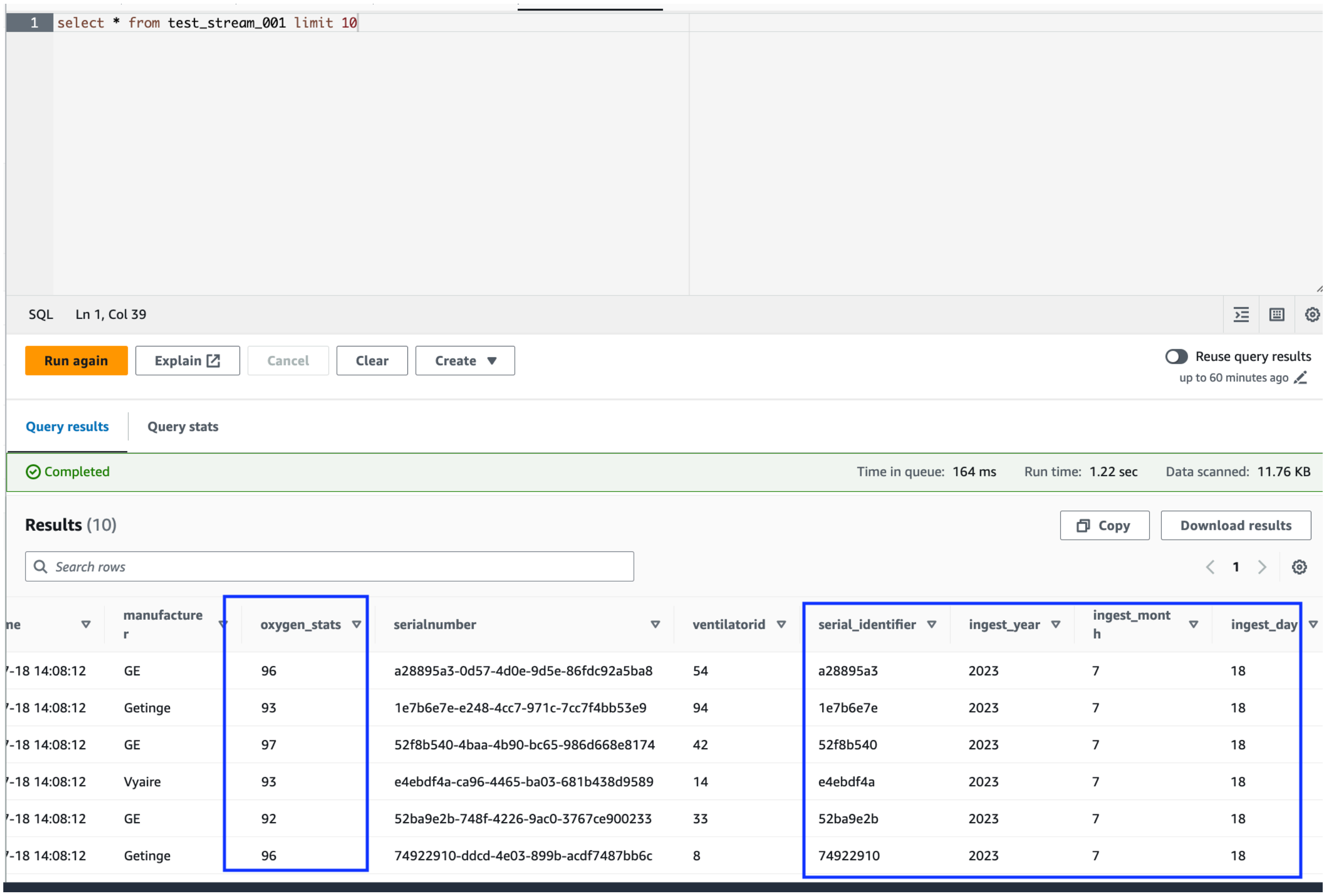
    b. Rename a column (o2stats to oxygen_stats).

    c. Derive new columns (serial_identifier, ingest_year, ingest_month and ingest_day).

    d. Store the results into an Amazon S3 bucket and also create a partitioned AWS Glue catalog table

8. In the last cell, you trigger the process batch every 10 seconds. Run the cell and wait for about 30 seconds for it to populate the Amazon S3 bucket and the AWS Glue catalog table.

9. Finally, browse the stored data using the Amazon Athena query editor. You can see the renamed column and also the new partitions.



```
select * from test_stream_001 limit 10
```

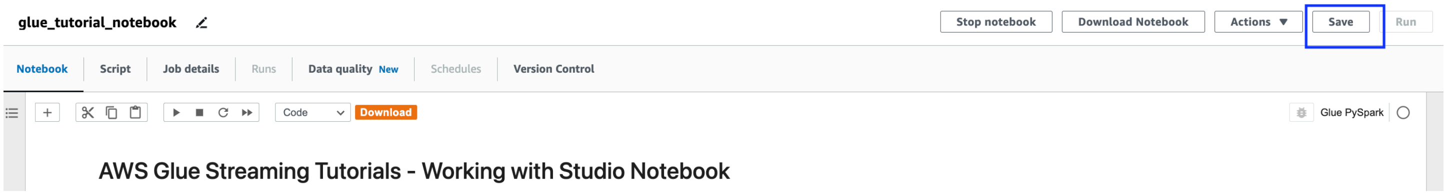
SQL
Ln 1, Col 39
Run again
Explain
Cancel
Clear
Create ▼
Reuse query results
up to 60 minutes ago
Query results
Query stats
Completed
Time in queue: 164 ms
Run time: 1.22 sec
Data scanned: 11.76 KB
Results (10)
Copy
Download results
Search rows
1

ne	manufacturer	oxygen_stats	serialnumber	ventilatorid	serial_identifier	ingest_year	ingest_month	ingest_day
-18 14:08:12	GE	96	a28895a3-0d57-4d0e-9d5e-86fdc92a5ba8	54	a28895a3	2023	7	18
-18 14:08:12	Getinge	93	1e7b6e7e-e248-4cc7-971c-7cc7f4bb53e9	94	1e7b6e7e	2023	7	18
-18 14:08:12	GE	97	52f8b540-4baa-4b90-bc65-986d668e8174	42	52f8b540	2023	7	18
-18 14:08:12	Vyaire	93	e4ebdf4a-ca96-4465-ba03-681b438d9589	14	e4ebdf4a	2023	7	18
-18 14:08:12	GE	92	52ba9e2b-748f-4226-9ac0-3767ce900233	33	52ba9e2b	2023	7	18
-18 14:08:12	Getinge	96	74922910-ddcd-4e03-899b-acdf7487bb6c	8	74922910	2023	7	18

The notebook has all necessary instructions to continue the tutorial. You can either run the instructions on the notebook or follow along with this tutorial to continue with the job development.

Save and run the AWS Glue job

With the development and testing of your application complete using the interactive sessions notebook, click **Save** at the top of the notebook interface. Once saved you can also run the application as a job.



glue_tutorial_notebook
Stop notebook
Download Notebook
Actions ▼
Save
Run
Notebook
Script
Job details
Runs
Data quality New
Schedules
Version Control
Code
Download
Glue PySpark

AWS Glue Streaming Tutorials - Working with Studio Notebook

Clean up

To avoid incurring additional charges to your account, stop the streaming job that you started as part of the instructions. You can do this by stopping the notebook, which will end the session. Empty the Amazon S3 bucket and delete the AWS CloudFormation stack that you provisioned earlier.

Conclusion

In this tutorial, we demonstrated how to do the following using the AWS Glue Studio notebook

- Author a streaming ETL job using notebooks
- Preview incoming data streams
- Code and fix issues without having to publish AWS Glue jobs
- Review the end-to-end working code, remove any debugging, and print statements or cells from the notebook
- Publish the code as an AWS Glue job

The goal of this tutorial is to give you hands-on experience working with AWS Glue Streaming and interactive sessions. We encourage you to use this as a reference for your individual AWS Glue Streaming use cases. For more information, see Getting started with AWS Glue interactive sessions.

AWS Glue Streaming concepts

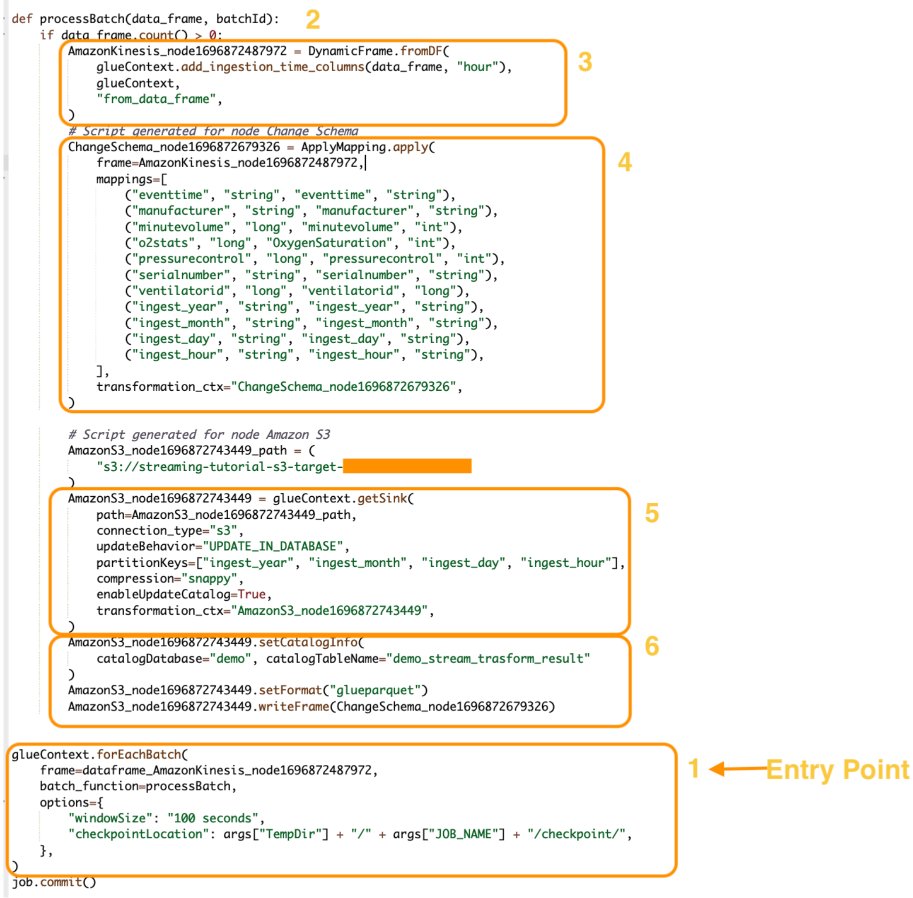
The following sections provide information on concepts of AWS Glue Streaming.

Topics

- Anatomy of a AWS Glue streaming job

Anatomy of a AWS Glue streaming job

AWS Glue streaming jobs operate on the Spark streaming paradigm and leverage structured streaming from the Spark framework. Streaming jobs constantly poll on the streaming data source, at a specific interval of time, to fetch records as micro batches. The following sections examine the different parts of a AWS Glue streaming job.



```
def processBatch(data_frame, batchId):
    if data_frame.count() > 0:
        AmazonKinesis_node1696872487972 = DynamicFrame.fromDF(
            glueContext.add_ingestion_time_columns(data_frame, "hour"),
            glueContext,
            "from_data_frame",
        )
        # Script generated for node Change Schema
        ChangeSchema_node1696872679326 = ApplyMapping.apply(
            frame=AmazonKinesis_node1696872487972,
            mappings=[
                ("eventtime", "string", "eventtime", "string"),
                ("manufacturer", "string", "manufacturer", "string"),
                ("minutevolume", "long", "minutevolume", "int"),
                ("o2stats", "long", "OxygenSaturation", "int"),
                ("pressurecontrol", "long", "pressurecontrol", "int"),
                ("serialnumber", "string", "serialnumber", "string"),
                ("ventilatorid", "long", "ventilatorid", "long"),
                ("ingest_year", "string", "ingest_year", "string"),
                ("ingest_month", "string", "ingest_month", "string"),
                ("ingest_day", "string", "ingest_day", "string"),
                ("ingest_hour", "string", "ingest_hour", "string"),
            ],
            transformation_ctx="ChangeSchema_node1696872679326",
        )

        # Script generated for node Amazon S3
        AmazonS3_node1696872743449_path = (
            "s3://streaming-tutorial-s3-target-[redacted]
        )
        AmazonS3_node1696872743449 = glueContext.getSink(
            path=AmazonS3_node1696872743449_path,
            connection_type="s3",
            updateBehavior="UPDATE_IN_DATABASE",
            partitionKeys=["ingest_year", "ingest_month", "ingest_day", "ingest_hour"],
            compression="snappy",
            enableUpdateCatalog=True,
            transformation_ctx="AmazonS3_node1696872743449",
        )
        AmazonS3_node1696872743449.setCatalogInfo(
            catalogDatabase="demo", catalogTableName="demo_stream_trasform_result"
        )
        AmazonS3_node1696872743449.setFormat("glueparquet")
        AmazonS3_node1696872743449.writeFrame(ChangeSchema_node1696872679326)

glueContext.forEachBatch(
    frame=dataframe_AmazonKinesis_node1696872487972,
    batch_function=processBatch,
    options={
        "windowSize": "100 seconds",
        "checkpointLocation": args["TempDir"] + "/" + args["JOB_NAME"] + "/checkpoint/",
    },
)
job.commit()
```

forEachBatch

The `forEachBatch` method is the entry point of a AWS Glue streaming job run. AWS Glue streaming jobs uses the `forEachBatch` method to poll data functioning like an iterator that remains active during the lifecycle of the streaming job and regularly polls the streaming source for new data and processes the latest data in micro batches.

```
glueContext.forEachBatch(
    frame=dataFrame_AmazonKinesis_node1696872487972,
    batch_function=processBatch,
    options={
        "windowSize": "100 seconds",
        "checkpointLocation": args["TempDir"] + "/" + args["JOB_NAME"] + "/
checkpoint/",
    },
)
```

Configure the `frame` property of `forEachBatch` to specify a streaming source. In this example, the source node that you created in the blank canvas during job creation is populated with the default DataFrame of the job. Set the `batch_function` property as the `function` that you decide to invoke for each micro batch operation. You must define a function to handle the batch transformation on the incoming data.

Source

In the first step of the `processBatch` function, the program verifies the record count of the DataFrame that you defined as frame property of `forEachBatch`. The program appends an ingestion time stamp to a non-empty DataFrame. The `data_frame.count()>0` clause determines whether the latest micro batch is not empty and is ready for further processing.

```
def processBatch(data_frame, batchId):
    if data_frame.count() >0:
        AmazonKinesis_node1696872487972 = DynamicFrame.fromDF(
            glueContext.add_ingestion_time_columns(data_frame, "hour"),
            glueContext,
            "from_data_frame",
        )
```

Mapping

The next section of the program is to apply mapping. The `Mapping.apply` method on a spark DataFrame allows you to define transformation rule around data elements. Typically you can rename, change the data type, or apply a custom function on the source data column and map those to the target columns.

```

    #Script generated for node ChangeSchema
    ChangeSchema_node16986872679326 = ApplyMapping.apply(
        frame =  AmazonKinesis_node1696872487972,
        mappings = [
            ("eventtime", "string", "eventtime", "string"),
            ("manufacturer", "string", "manufacturer", "string"),
            ("minutevolume", "long", "minutevolume", "int"),
            ("o2stats", "long", "OxygenSaturation", "int"),
            ("pressurecontrol", "long", "pressurecontrol", "int"),
            ("serialnumber", "string", "serialnumber", "string"),
            ("ventilatorid", "long", "ventilatorid", "long"),
            ("ingest_year", "string", "ingest_year", "string"),
            ("ingest_month", "string", "ingest_month", "string"),
            ("ingest_day", "string", "ingest_day", "string"),
            ("ingest_hour", "string", "ingest_hour", "string"),
        ],
        transformation_ctx="ChangeSchema_node16986872679326",
    )
        )
```

Sink

In this section, the incoming data set from the streaming source are stored at a target location. In this example we will write the data to an Amazon S3 location. The `AmazonS3_node_path` property details is pre-populated as determined by the settings you used during job creation from the canvas. You can set the `updateBehavior` based on your use case and decide to either Not update the data catalog table, or Create data catalog and update data catalog schema on subsequent runs, or create a catalog table and not update the schema definition on subsequent runs.

The `partitionKeys` property defines the storage partition option. The default behavior is to partition the data per the `ingestion_time_columns` that was made available in the source section. The `compression` property allows you to set the compression algorithm to be applied during target write. You have options to set Snappy, LZO, or GZIP as the compression technique. The `enableUpdateCatalog` property controls whether the AWS Glue catalog table needs to be updated. Available options for this property are `True` or `False`.

```
    #Script generated for node Amazon S3
    AmazonS3_node1696872743449 = glueContext.getSink(
        path =  AmazonS3_node1696872743449_path,
        connection_type = "s3",
        updateBehavior = "UPDATE_IN_DATABASE",
        partitionKeys = ["ingest_year", "ingest_month", "ingest_day", "ingest_hour"],
        compression = "snappy",
        enableUpdateCatalog = True,
        transformation_ctx = "AmazonS3_node1696872743449",
    )
```

AWS Glue Catalog sink

This section of the job controls the AWS Glue catalog table update behavior. Set `catalogDatabase` and `catalogTableName` property per your AWS Glue Catalog database name and the table name associated with the AWS Glue job that you are designing. You can define the file format of the target data via the `setFormat` property. For this example we will store the data in parquet format.

Once you set up and run the AWS Glue streaming job referring this tutorial, the streaming data produced at Amazon Kinesis Data Streams will be stored at the Amazon S3 location in a parquet format with snappy compression. On successful runs of the streaming job you will able to query the data through Amazon Athena.

```
AmazonS3_node1696872743449 = setCatalogInfo(
    catalogDatabase =  "demo", catalogTableName = "demo_stream_transform_result"
)
AmazonS3_node1696872743449.setFormat("glueparquet")
AmazonS3_node1696872743449.writeFormat("ChangeSchema_node16986872679326")
)
```

AWS Glue Streaming connections

The following sections provide information on how to use connections in AWS Glue Streaming.

Topics

- Working with Kafka connections
- Working with Kinesis connections

Working with Kafka connections

You can use a Kafka connection to read and write to Kafka data streams using information stored in a Data Catalog table, or by providing information to directly access the data stream. The connection supports a Kafka cluster or an Amazon Managed Streaming for Apache Kafka cluster. You can read information from Kafka into a Spark DataFrame, then convert it to a AWS Glue

DynamicFrame. You can write DynamicFrames to Kafka in a JSON format. If you directly access the data stream, use these options to provide the information about how to access the data stream.

If you use `getCatalogSource` or `create_data_frame_from_catalog` to consume records from a Kafka streaming source, or `getCatalogSink` or `write_dynamic_frame_from_catalog` to write records to Kafka, and the job has the Data Catalog database and table name information, and can use that to obtain some basic parameters for reading from the Kafka streaming source. If you use `getSource`, `getCatalogSink`, `getSourceWithFormat`, `getSinkWithFormat`, `createDataFrameFromOptions` or `create_data_frame_from_options`, or `write_dynamic_frame_from_catalog`, you must specify these basic parameters using the connection options described here.

You can specify the connection options for Kafka using the following arguments for the specified methods in the `GlueContext` class.

- Scala
  - `connectionOptions`: Use with `getSource`, `createDataFrameFromOptions`, `getSink`
  - `additionalOptions`: Use with `getCatalogSource`, `getCatalogSink`
  - `options`: Use with `getSourceWithFormat`, `getSinkWithFormat`
- Python
  - `connection_options`: Use with `create_data_frame_from_options`, `write_dynamic_frame_from_options`
  - `additional_options`: Use with `create_data_frame_from_catalog`, `write_dynamic_frame_from_catalog`
  - `options`: Use with `getSource`, `getSink`

For notes and restrictions about streaming ETL jobs, consult the section called “Streaming ETL notes and restrictions”.

Configure Kafka

There are no AWS prerequisites to connecting to Kafka streams available through the internet.

You can create a AWS Glue Kafka connection to manage your connection credentials. For more information, see the section called “Creating a connection for a Kafka data stream”. In your AWS Glue job configuration, provide *`connectionName`* as an **Additional network connection**, then, in your method call, provide *`connectionName`* to the `connectionName` parameter.

In certain cases, you will need to configure additional prerequisites:

- If using Amazon Managed Streaming for Apache Kafka with IAM authentication, you will need appropriate IAM configuration.
- If using Amazon Managed Streaming for Apache Kafka within an Amazon VPC, you will need appropriate Amazon VPC configuration. You will need to create a AWS Glue connection that provides Amazon VPC connection information. You will need your job configuration to include the AWS Glue connection as an **Additional network connection**.

For more information about Streaming ETL job prerequisites, consult the section called "Streaming ETL jobs".

Example: Reading from Kafka streams

Used in conjunction with the section called "forEachBatch".

Example for Kafka streaming source:

```
kafka_options =
    { "connectionName": "ConfluentKafka",
      "topicName": "kafka-auth-topic",
      "startingOffsets": "earliest",
      "inferSchema": "true",
      "classification": "json"
    }
data_frame_datasource0 =
 glueContext.create_data_frame.from_options(connection_type="kafka",
 connection_options=kafka_options)
```

Example: Writing to Kafka streams

Examples for writing to Kafka:

Example with the `getSink` method:

```
data_frame_datasource0 =
glueContext.getSink(
 connectionType="kafka",
 connectionOptions={
  JsonOptions("""{
   "connectionName": "ConfluentKafka",
    "classification": "json",
    "topic": "kafka-auth-topic",
    "typeOfData": "kafka"}
  """)},
 transformationContext="dataframe_ApacheKafka_node1711729173428")
 .getDataFrame()
```

Example with the `write_dynamic_frame.from_options` method:

```
kafka_options =
    { "connectionName": "ConfluentKafka",
      "topicName": "kafka-auth-topic",
      "classification": "json"
    }
data_frame_datasource0 =
 glueContext.write_dynamic_frame.from_options(connection_type="kafka",
 connection_options=kafka_options)
```

Kafka connection option reference

When reading, use the following connection options with `"connectionType": "kafka"`:

- `"bootstrap.servers"` (Required) A list of bootstrap server URLs, for example, as `b-1.vpc-test-2.o4q88o.c6.kafka.us-east-1.amazonaws.com:9094`. This option must be specified in the API call or defined in the table metadata in the Data Catalog.
- `"security.protocol"` (Required) The protocol used to communicate with brokers. The possible values are `"SSL"` or `"PLAINTEXT"`.
- `"topicName"` (Required) A comma-separated list of topics to subscribe to. You must specify one and only one of `"topicName"`, `"assign"` or `"subscribePattern"`.
- `"assign"`: (Required) A JSON string specifying the specific `TopicPartitions` to consume. You must specify one and only one of `"topicName"`, `"assign"` or `"subscribePattern"`.

  Example: '{"topicA":[0,1],"topicB":[2,4]}'
- `"subscribePattern"`: (Required) A Java regex string that identifies the topic list to subscribe to. You must specify one and only one of `"topicName"`, `"assign"` or `"subscribePattern"`.

  Example: 'topic.*'
- `"classification"` (Required) The file format used by the data in the record. Required unless provided through the Data Catalog.

- `"delimiter"` (Optional) The value separator used when `classification` is CSV. Default is "`,`."
- `"startingOffsets"`: (Optional) The starting position in the Kafka topic to read data from. The possible values are `"earliest"` or `"latest"`. The default value is `"latest"`.
- `"startingTimestamp"`: (Optional, supported only for AWS Glue version 4.0 or later) The Timestamp of the record in the Kafka topic to read data from. The possible value is a Timestamp string in UTC format in the pattern `yyyy-mm-ddTHH:MM:SSZ` (where `Z` represents a UTC timezone offset with a +/-. For example: "2023-04-04T08:00:00-04:00").

  Note: Only one of 'startingOffsets' or 'startingTimestamp' can be present in the Connection Options list of the AWS Glue streaming script, including both these properties will result in job failure.
- `"endingOffsets"`: (Optional) The end point when a batch query is ended. Possible values are either `"latest"` or a JSON string that specifies an ending offset for each `TopicPartition`.

  For the JSON string, the format is `{"topicA":{"0":23,"1":-1},"topicB":{"0":-1}}`. The value `-1` as an offset represents `"latest"`.
- `"pollTimeoutMs"`: (Optional) The timeout in milliseconds to poll data from Kafka in Spark job executors. The default value is `600000`.
- `"numRetries"`: (Optional) The number of times to retry before failing to fetch Kafka offsets. The default value is `3`.
- `"retryIntervalMs"`: (Optional) The time in milliseconds to wait before retrying to fetch Kafka offsets. The default value is `10`.
- `"maxOffsetsPerTrigger"`: (Optional) The rate limit on the maximum number of offsets that are processed per trigger interval. The specified total number of offsets is proportionally split across `topicPartitions` of different volumes. The default value is null, which means that the consumer reads all offsets until the known latest offset.
- `"minPartitions"`: (Optional) The desired minimum number of partitions to read from Kafka. The default value is null, which means that the number of spark partitions is equal to the number of Kafka partitions.
- `"includeHeaders"`: (Optional) Whether to include the Kafka headers. When the option is set to "true", the data output will contain an additional column named "glue_streaming_kafka_headers" with type `Array[Struct(key: String, value: String)]`. The default value is "false". This option is available in AWS Glue version 3.0 or later.

- `"schema"`: (Required when inferSchema set to false) The schema to use to process the payload. If classification is `avro` the provided schema must be in the Avro schema format. If the classification is not `avro` the provided schema must be in the DDL schema format.

  The following are schema examples.

  Example in DDL schema format

  ```
  'column1' INT, 'column2' STRING , 'column3' FLOAT
  ```

  Example in Avro schema format

  ```
  {
  "type":"array",
  "items":
  {
  "type":"record",
  "name":"test",
  "fields":
  [
    {
      "name":"_id",
      "type":"string"
    },
    {
      "name":"index",
      "type":
      [
        "int",
        "string",
        "float"
      ]
    }
  ]
  }
  }
  ```
- `"inferSchema"`: (Optional) The default value is 'false'. If set to 'true', the schema will be detected at runtime from the payload within `foreachbatch`.
- `"avroSchema"`: (Deprecated) Parameter used to specify a schema of Avro data when Avro format is used. This parameter is now deprecated. Use the `schema` parameter.

- `"addRecordTimestamp"`: (Optional) When this option is set to 'true', the data output will contain an additional column named "__src_timestamp" that indicates the time when the corresponding record received by the topic. The default value is 'false'. This option is supported in AWS Glue version 4.0 or later.
- `"emitConsumerLagMetrics"`: (Optional) When the option is set to 'true', for each batch, it will emit the metrics for the duration between the oldest record received by the topic and the time it arrives in AWS Glue to CloudWatch. The metric's name is "glue.driver.streaming.maxConsumerLagInMs". The default value is 'false'. This option is supported in AWS Glue version 4.0 or later.

When writing, use the following connection options with `"connectionType": "kafka"`:

- `"connectionName"` (Required) Name of the AWS Glue connection used to connect to the Kafka cluster (similar to Kafka source).
- `"topic"` (Required) If a topic column exists then its value is used as the topic when writing the given row to Kafka, unless the topic configuration option is set. That is, the `topic` configuration option overrides the topic column.
- `"partition"` (Optional) If a valid partition number is specified, that `partition` will be used when sending the record.

  If no partition is specified but a `key` is present, a partition will be chosen using a hash of the key.

  If neither `key` nor `partition` is present, a partition will be chosen based on sticky partitioning those changes when at least batch.size bytes are produced to the partition.
- `"key"` (Optional) Used for partitioning if `partition` is null.
- `"classification"` (Optional) The file format used by the data in the record. We only support JSON, CSV and Avro.

  With Avro format, we can provide a custom avroSchema to serialize with, but note that this needs to be provided on the source for deserializing as well. Else, by default it uses the Apache AvroSchema for serializing.

Additionally, you can fine-tune the Kafka sink as required by updating the [Kafka producer configuration parameters](). Note that there is no allow listing on connection options, all the key-value pairs are persisted on the sink as is.

However, there is a small deny list of options that will not take effect. For more information, see Kafka specific configurations.

Working with Kinesis connections

You can use a Kinesis connection to read and write to Amazon Kinesis data streams using information stored in a Data Catalog table, or by providing information to directly access the data stream. You can read information from Kinesis into a Spark DataFrame, then convert it to a AWS Glue DynamicFrame. You can write DynamicFrames to Kinesis in a JSON format. If you directly access the data stream, use these options to provide the information about how to access the data stream.

If you use `getCatalogSource` or `create_data_frame_from_catalog` to consume records from a Kinesis streaming source, the job has the Data Catalog database and table name information, and can use that to obtain some basic parameters for reading from the Kinesis streaming source. If you use `getSource`, `getSourceWithFormat`, `createDataFrameFromOptions` or `create_data_frame_from_options`, you must specify these basic parameters using the connection options described here.

You can specify the connection options for Kinesis using the following arguments for the specified methods in the `GlueContext` class.

- Scala
  - `connectionOptions`: Use with `getSource`, `createDataFrameFromOptions`, `getSink`
  - `additionalOptions`: Use with `getCatalogSource`, `getCatalogSink`
  - `options`: Use with `getSourceWithFormat`, `getSinkWithFormat`
- Python
  - `connection_options`: Use with `create_data_frame_from_options`, `write_dynamic_frame_from_options`
  - `additional_options`: Use with `create_data_frame_from_catalog`, `write_dynamic_frame_from_catalog`
  - `options`: Use with `getSource`, `getSink`

For notes and restrictions about Streaming ETL jobs, consult the section called “Streaming ETL notes and restrictions”.

Configure Kinesis

To connect to a Kinesis data stream in an AWS Glue Spark job, you will need some prerequisites:

- If reading, the AWS Glue job must have Read access level IAM permissions to the Kinesis data stream.
- If writing, the AWS Glue job must have Write access level IAM permissions to the Kinesis data stream.

In certain cases, you will need to configure additional prerequisites:

- If your AWS Glue job is configured with **Additional network connections** (typically to connect to other datasets) and one of those connections provides Amazon VPC **Network options**, this will direct your job to communicate over Amazon VPC. In this case you will also need to configure your Kinesis data stream to communicate over Amazon VPC. You can do this by creating an interface VPC endpoint between your Amazon VPC and Kinesis data stream. For more information, see Using Kinesis Data Streams with Interface VPC Endpoints.
- When specifying Amazon Kinesis Data Streams in another account, you must setup the roles and policies to allow cross-account access. For more information, see Example: Read From a Kinesis Stream in a Different Account.

For more information about Streaming ETL job prerequisites, consult the section called “Streaming ETL jobs”.

Read from Kinesis

Example: Reading from Kinesis streams

Used in conjunction with the section called “forEachBatch”.

Example for Amazon Kinesis streaming source:

```
kinesis_options =
   { "streamARN": "arn:aws:kinesis:us-east-2:777788889999:stream/fromOptionsStream",
     "startingPosition": "TRIM_HORIZON",
     "inferSchema": "true",
     "classification": "json"
   }
data_frame_datasource0 =
 glueContext.create_data_frame.from_options(connection_type="kinesis",
 connection_options=kinesis_options)
```

Write to Kinesis

Example: Writing to Kinesis streams

Used in conjunction with the section called “forEachBatch”. Your DynamicFrame will be written to the stream in a JSON format. If the job cannot write after several retries, it will fail. By default, each DynamicFrame record will be sent to the Kinesis stream individually. You can configure this behavior using `aggregationEnabled` and associated parameters.

Example writing to Amazon Kinesis from a streaming job:

Python

```
glueContext.write_dynamic_frame.from_options(
    frame=frameToWrite
    connection_type="kinesis",
    connection_options={
        "partitionKey": "part1",
        "streamARN": "arn:aws:kinesis:us-east-1:111122223333:stream/streamName",
    }
)
```

Scala

```
glueContext.getSinkWithFormat(
                connectionType="kinesis",
                options=JsonOptions("""{
                    "streamARN": "arn:aws:kinesis:us-
east-1:111122223333:stream/streamName",
                    "partitionKey": "part1"
                }"""),
           )
           .writeDynamicFrame(frameToWrite)
```

Kinesis connection parameters

Designates connection options for Amazon Kinesis Data Streams.

Use the following connection options for Kinesis streaming data sources:

- `"streamARN"` (Required) Used for Read/Write. The ARN of the Kinesis data stream.

- `"classification"` (Required for read) Used for Read. The file format used by the data in the record. Required unless provided through the Data Catalog.
- `"streamName"` – (Optional) Used for Read. The name of a Kinesis data stream to read from. Used with `endpointUrl`.
- `"endpointUrl"` – (Optional) Used for Read. Default: "https://kinesis.us-east-1.amazonaws.com". The AWS endpoint of the Kinesis stream. You do not need to change this unless you are connecting to a special region.
- `"partitionKey"` – (Optional) Used for Write. The Kinesis partition key used when producing records.
- `"delimiter"` (Optional) Used for Read. The value separator used when `classification` is CSV. Default is "`,`."
- `"startingPosition"`: (Optional) Used for Read. The starting position in the Kinesis data stream to read data from. The possible values are `"latest"`, `"trim_horizon"`, `"earliest"`, or a Timestamp string in UTC format in the pattern `yyyy-mm-ddTHH:MM:SSZ` (where `Z` represents a UTC timezone offset with a +/-. For example "2023-04-04T08:00:00-04:00"). The default value is `"latest"`. Note: the Timestamp string in UTC Format for `"startingPosition"` is supported only for AWS Glue version 4.0 or later.
- `"failOnDataLoss"`: (Optional) Fail the job if any active shard is missing or expired. The default value is `"false"`.
- `"awsSTSRoleARN"`: (Optional) Used for Read/Write. The Amazon Resource Name (ARN) of the role to assume using AWS Security Token Service (AWS STS). This role must have permissions for describe or read record operations for the Kinesis data stream. You must use this parameter when accessing a data stream in a different account. Used in conjunction with `"awsSTSSessionName"`.
- `"awsSTSSessionName"`: (Optional) Used for Read/Write. An identifier for the session assuming the role using AWS STS. You must use this parameter when accessing a data stream in a different account. Used in conjunction with `"awsSTSRoleARN"`.
- `"awsSTSEndpoint"`: (Optional) The AWS STS endpoint to use when connecting to Kinesis with an assumed role. This allows using the regional AWS STS endpoint in a VPC, which is not possible with the default global endpoint.
- `"maxFetchTimeInMs"`: (Optional) Used for Read. The maximum time spent for the job executor to read records for the current batch from the Kinesis data stream, specified in milliseconds (ms). Multiple `GetRecords` API calls may be made within this time. The default value is `1000`.

- `"maxFetchRecordsPerShard"`: (Optional) Used for Read. The maximum number of records to fetch per shard in the Kinesis data stream per microbatch. Note: The client can exceed this limit if the streaming job has already read extra records from Kinesis (in the same get-records call). If `maxFetchRecordsPerShard` needs to be strict then it needs to be a multiple of `maxRecordPerRead`. The default value is `100000`.
- `"maxRecordPerRead"`: (Optional) Used for Read. The maximum number of records to fetch from the Kinesis data stream in each `getRecords` operation. The default value is `10000`.
- `"addIdleTimeBetweenReads"`: (Optional) Used for Read. Adds a time delay between two consecutive `getRecords` operations. The default value is `"False"`. This option is only configurable for Glue version 2.0 and above.
- `"idleTimeBetweenReadsInMs"`: (Optional) Used for Read. The minimum time delay between two consecutive `getRecords` operations, specified in ms. The default value is `1000`. This option is only configurable for Glue version 2.0 and above.
- `"describeShardInterval"`: (Optional) Used for Read. The minimum time interval between two `ListShards` API calls for your script to consider resharding. For more information, see Strategies for Resharding in *Amazon Kinesis Data Streams Developer Guide*. The default value is `1s`.
- `"numRetries"`: (Optional) Used for Read. The maximum number of retries for Kinesis Data Streams API requests. The default value is `3`.
- `"retryIntervalMs"`: (Optional) Used for Read. The cool-off time period (specified in ms) before retrying the Kinesis Data Streams API call. The default value is `1000`.
- `"maxRetryIntervalMs"`: (Optional) Used for Read. The maximum cool-off time period (specified in ms) between two retries of a Kinesis Data Streams API call. The default value is `10000`.
- `"avoidEmptyBatches"`: (Optional) Used for Read. Avoids creating an empty microbatch job by checking for unread data in the Kinesis data stream before the batch is started. The default value is `"False"`.
- `"schema"`: (Required when inferSchema set to false) Used for Read. The schema to use to process the payload. If classification is `avro` the provided schema must be in the Avro schema format. If the classification is not `avro` the provided schema must be in the DDL schema format.

  The following are schema examples.

Example in DDL schema format

```
`column1` INT, `column2` STRING , `column3` FLOAT
```

Example in Avro schema format

```
{
  "type":"array",
  "items":
  {
    "type":"record",
    "name":"test",
    "fields":
    [
      {
        "name":"_id",
        "type":"string"
      },
      {
        "name":"index",
        "type":
        [
          "int",
          "string",
          "float"
        ]
      }
    ]
  }
}
```

- `"inferSchema"`: (Optional) Used for Read. The default value is 'false'. If set to 'true', the schema will be detected at runtime from the payload within `foreachbatch`.
- `"avroSchema"`: (Deprecated) Used for Read. Parameter used to specify a schema of Avro data when Avro format is used. This parameter is now deprecated. Use the `schema` parameter.
- `"addRecordTimestamp"`: (Optional) Used for Read. When this option is set to 'true', the data output will contain an additional column named "__src_timestamp" that indicates the time when the corresponding record received by the stream. The default value is 'false'. This option is supported in AWS Glue version 4.0 or later.

- `"emitConsumerLagMetrics"`: (Optional) Used for Read. When the option is set to 'true', for each batch, it will emit the metrics for the duration between the oldest record received by the stream and the time it arrives in AWS Glue to CloudWatch. The metric's name is "glue.driver.streaming.maxConsumerLagInMs". The default value is 'false'. This option is supported in AWS Glue version 4.0 or later.
- `"fanoutConsumerARN"`: (Optional) Used for Read. The ARN of a Kinesis stream consumer for the stream specified in `streamARN`. Used to enable enhanced fan-out mode for your Kinesis connection. For more information on consuming a Kinesis stream with enhanced fan-out, see the section called "Using enhanced fan-out in Kinesis streaming jobs".
- `"recordMaxBufferedTime"` – (Optional) Used for Write. Default: 1000 (ms). Maximum time a record is buffered while waiting to be written.
- `"aggregationEnabled"` – (Optional) Used for Write. Default: true. Specifies if records should be aggregated before sending them to Kinesis.
- `"aggregationMaxSize"` – (Optional) Used for Write. Default: 51200 (bytes). If a record is larger than this limit, it will bypass the aggregator. Note Kinesis enforces a limit of 50KB on record size. If you set this beyond 50KB, oversize records will be rejected by Kinesis.
- `"aggregationMaxCount"` – (Optional) Used for Write. Default: 4294967295. Maximum number of items to pack into an aggregated record.
- `"producerRateLimit"` – (Optional) Used for Write. Default: 150 (%). Limits per-shard throughput sent from a single producer (such as your job), as a percentage of the backend limit.
- `"collectionMaxCount"` – (Optional) Used for Write. Default: 500. Maximum number of items to pack into an PutRecords request.
- `"collectionMaxSize"` – (Optional) Used for Write. Default: 5242880 (bytes). Maximum amount of data to send with a PutRecords request.

AWS Glue streaming autoscaling

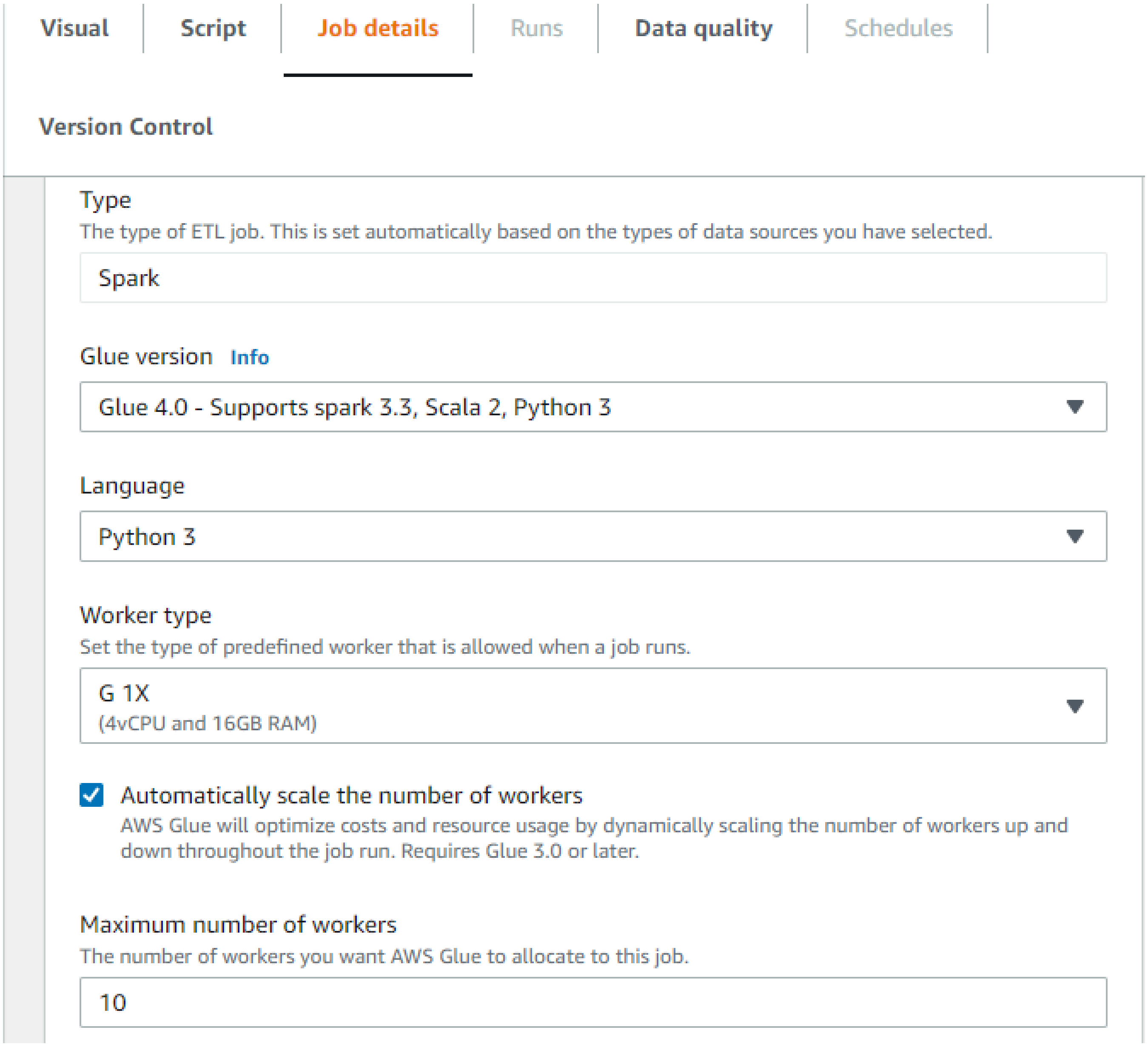
AWS Glue streaming ETL jobs continuously consume data from streaming sources, clean and transform the data in-flight, and make it available for analysis. By monitoring each stage of the job run, AWS Glue autoscaling can turn off workers when they are idle or add workers if additional parallel processing is possible.

The following sections provide information on AWS Glue streaming autoscaling

Enabling Auto Scaling in AWS Glue Studio

On the **Job details** tab in AWS Glue Studio, choose the type as **Spark** or **Spark Streaming**, and **Glue version** as **`Glue 3.0`** or **`Glue 4.0`**. Then a check box will show up below **Worker type**.

- Select the **Automatically scale the number of workers** option.
- Set the **Maximum number of workers** to define the maximum number of workers that can be vended to the job run.



Type
The type of ETL job. This is set automatically based on the types of data sources you have selected.
Spark

Glue version
Glue 4.0 - Supports spark 3.3, Scala 2, Python 3

Language
Python 3

Worker type
Set the type of predefined worker that is allowed when a job runs.
G 1X
(4vCPU and 16GB RAM)

☑ Automatically scale the number of workers
AWS Glue will optimize costs and resource usage by dynamically scaling the number of workers up and down throughout the job run. Requires Glue 3.0 or later.

Maximum number of workers
The number of workers you want AWS Glue to allocate to this job.
10

Enabling Auto Scaling with the AWS CLI or SDK

To enable Auto Scaling From the AWS CLI for your job run, run `start-job-run` with the following configuration:

```
{
    "JobName": "<your job name>",
    "Arguments": {
        "--enable-auto-scaling": "true"
    },
    "WorkerType": "G.2X", // G.1X, G.2X, G.4X, G.8X, G.12X, G.16X, R.1X, R.2X, R.4X,
 and R.8X are supported for Auto Scaling Jobs
    "NumberOfWorkers": 20, // represents Maximum number of workers
    ...other job run configurations...
}
```

Once at ETL job run is finished, you can also call `get-job-run` to check the actual resource usage of the job run in DPU-seconds. Note: the new field **DPUSeconds** will only show up for your batch jobs on AWS Glue 3.0 or later enabled with Auto Scaling. This field is not supported for streaming jobs.

```
$ aws glue get-job-run --job-name your-job-name --run-id jr_xx --endpoint https://
glue.us-east-1.amazonaws.com --region us-east-1
{
    "JobRun": {
        ...
        "GlueVersion": "3.0",
        "DPUSeconds": 386.0
    }
}
```

You can also configure job runs with Auto Scaling using the AWS Glue SDK with the same configuration.

How it works

Scaling across microbatch

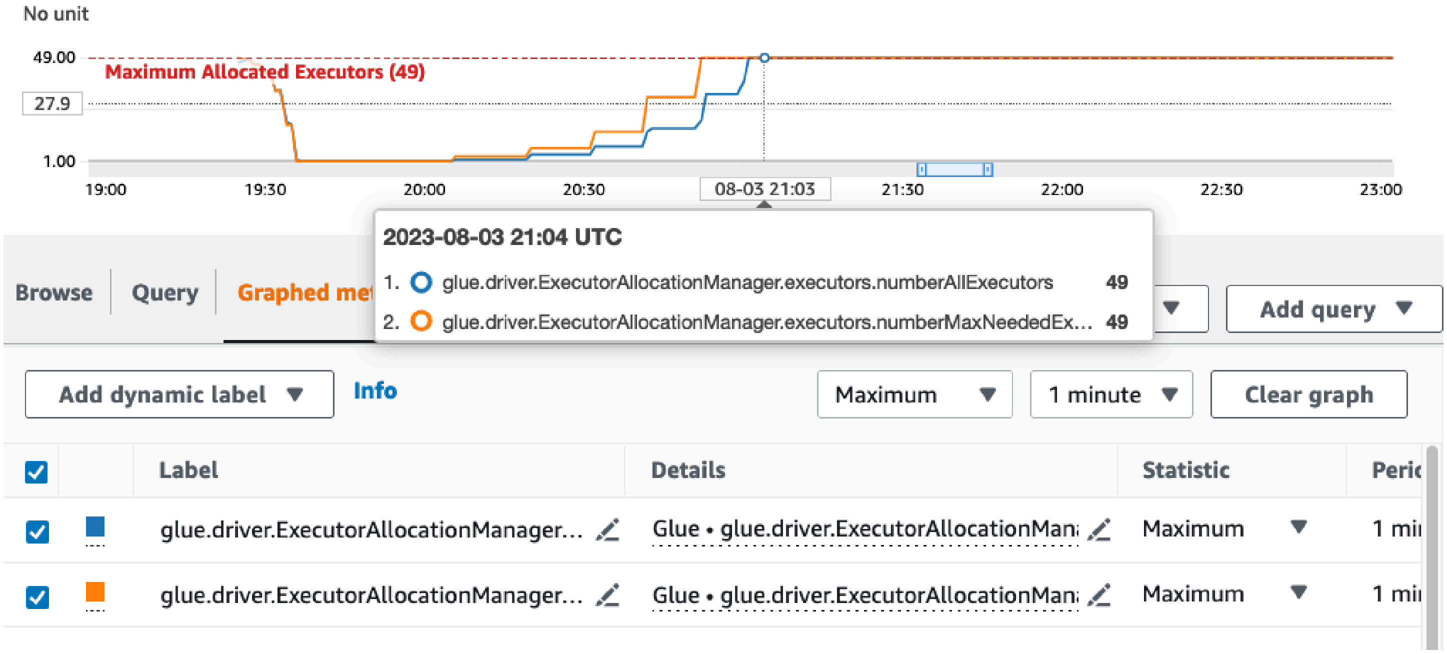
The following example is used to describe how autoscaling works.

- You have a AWS Glue job that starts with 50 DPUs.

- Autoscaling is enabled.

In this example, AWS Glue looks at the “batchProcessingTimeInMs“ metric for a few micro batches and determines if your jobs are completing within the window size that you have established. If your jobs are completing sooner and depending on how soon they complete, AWS Glue may scale down. This metric, plotted with ”numberAllExecutors“ can be monitored in Amazon CloudWatch to see how autoscaling works.

The number of executors exponentially scales up or down only after each micro batch completes. As you can see from the Amazon CloudWatch Monitoring log, AWS Glue looks at the number of needed executors (Orange Line) and scales the executors (blue line) to match that automatically.



2023-08-03 21:04 UTC

1. ○ glue.driver.ExecutorAllocationManager.executors.numberAllExecutors 49
2. ○ glue.driver.ExecutorAllocationManager.executors.numberMaxNeededEx... 49

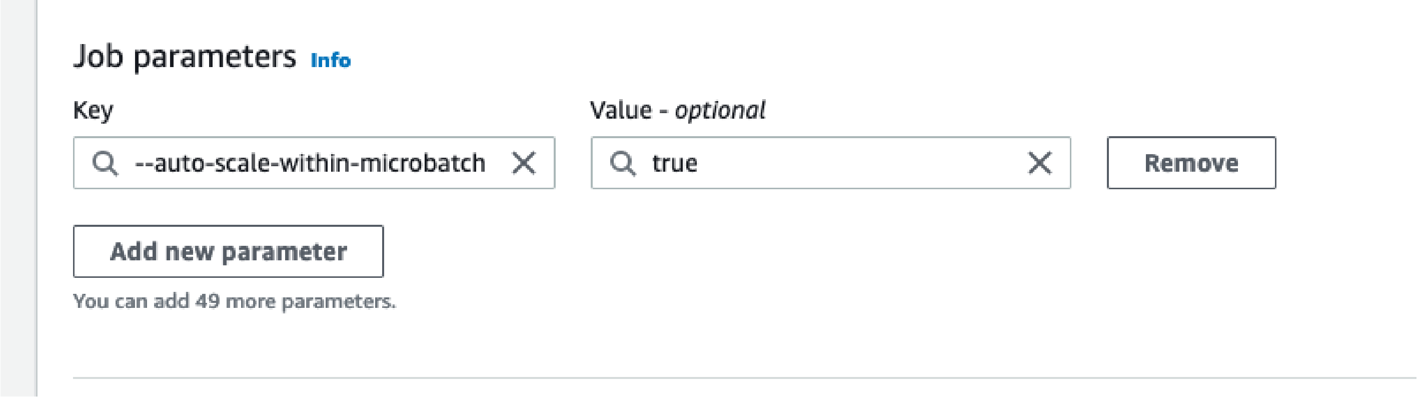
Maximum ▼ 1 minute ▼

☑		Label	Details	Statistic	Peric
☑	🟦	glue.driver.ExecutorAllocationManager...	Glue • glue.driver.ExecutorAllocationMan	Maximum ▼	1 mi
☑	🟧	glue.driver.ExecutorAllocationManager...	Glue • glue.driver.ExecutorAllocationMan	Maximum ▼	1 mi

Once AWS Glue scales down the number of executors and observes that data volumes increase, consequently increasing the micro batch processing time, AWS Glue will scale up to 50 DPUs, which is the specified upper limit.

Scaling within microbatch

In the above example, the system monitors a few completed micro-batches to make a decision on whether to scale up or down. Longer windows require autoscaling to respond more quickly within the microbatch, rather than waiting for a few micro batches. For these cases, you can use an additional configuration `--auto-scaIe-within-microbatch` to `true`. You can add this to the AWS Glue job properties in AWS Glue Studio as shown below.



Job parameters

Key: --auto-scale-within-microbatch

Value - *optional*: true

Maintenance windows for AWS Glue Streaming

AWS Glue periodically performs maintenance activities. During these maintenance windows, AWS Glue will need to restart your streaming jobs. You can control when the jobs are restarted by specifying maintenance windows. In this section, we outline where you can setup the maintenance window and specific behaviors you should consider.

Topics

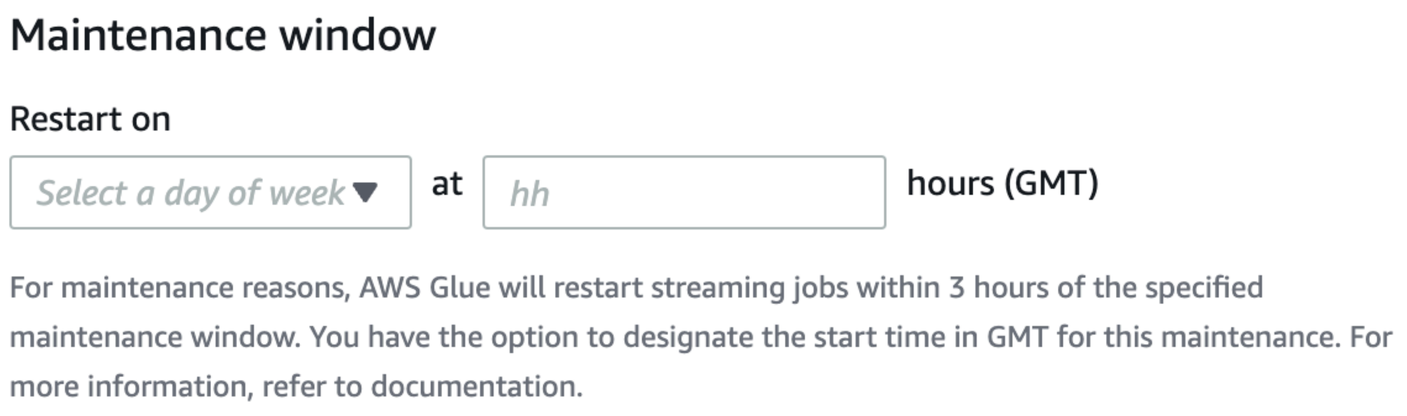
- Setting up a maintenance window
- Maintenance window behavior
- Job monitoring
- Data loss handling

Setting up a maintenance window

You can set up a maintenance window using AWS Glue Studio or APIs.

Setting up a maintenance windows in AWS Glue Studio

You can specify a maintenance window in the **Job Details** page of your AWS Glue Streaming job. You can specify the day and time in GMT. AWS Glue will restart your job within the specified time window.



Maintenance window

Restart on

For maintenance reasons, AWS Glue will restart streaming jobs within 3 hours of the specified maintenance window. You have the option to designate the start time in GMT for this maintenance. For more information, refer to documentation.

Setting up a maintenance windows in the API

You can alternatively set up the maintenance window in the Create Job API. Here is an example of configuring a maintenance windows via the API.

```
aws glue create-job –name jobName –role roleArnForTheJob –command
 Name=gluestreaming,ScriptLocation=s3-path-to-the-script --maintenance-window="Sun:10"
```

An example command is as follows:

```
aws glue create-job –name testMaintenance –role arn:aws:iam::012345678901:role/
Glue_DefaultRole –command Name=gluestreaming,ScriptLocation=s3://glue-example-test/
example.py –maintenance-window="Sun:10
```

Maintenance window behavior

AWS Glue goes through a series of steps to decide when to restart a job:

1. When a new streaming job is initiated, AWS Glue first checks if there is a timeout associated with the job run. A timeout allows you to configure the end time of the job. If the timeout is less than 7 days, then the job will not be restarted.
2. If the timeout is greater than 7 days, then AWS Glue checks if the maintenance window is configured for the job. If it is then that window is picked up and the window gets assigned to the job run. AWS Glue will restart the job within 3 hours of the specified maintenance window. For instance, if you set up the maintenance window for Monday at 10:00AM GMT, your jobs will be restarted between 10:00AM GMT to 1:00PM GMT.

3. If the maintenance window is not configured, AWS Glue automatically sets the restart time to 7 days past job run initiation time. For instance, if you initiated your job on 7/1/2024 12:00AM GMT and you did not specify maintenance windows, your job will be set to restart on 7/8/2024 at 12:00 AM GMT.

> **Note**
>
> If you are already running streaming jobs, this change will impact you starting July 1, 2024. You will have time until June 30th to configure your maintenance windows. After July 1st, any streaming jobs that you start will be restarted per this documentation. If you require any additional support, you can reach out to AWS Support.

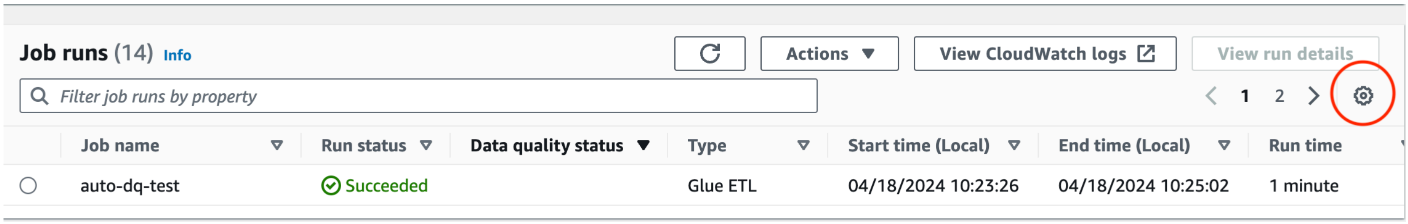
4. Sometimes, AWS Glue may not be able to restart the job, especially when the ongoing micro-batch is not processed. In these instances, the job will not be interrupted. In these instances, AWS Glue will restart the job after 14 days, and in this case, the maintenance window is not honored.

Job monitoring

You can monitor the jobs in the AWS Glue Studio **Monitoring** page.

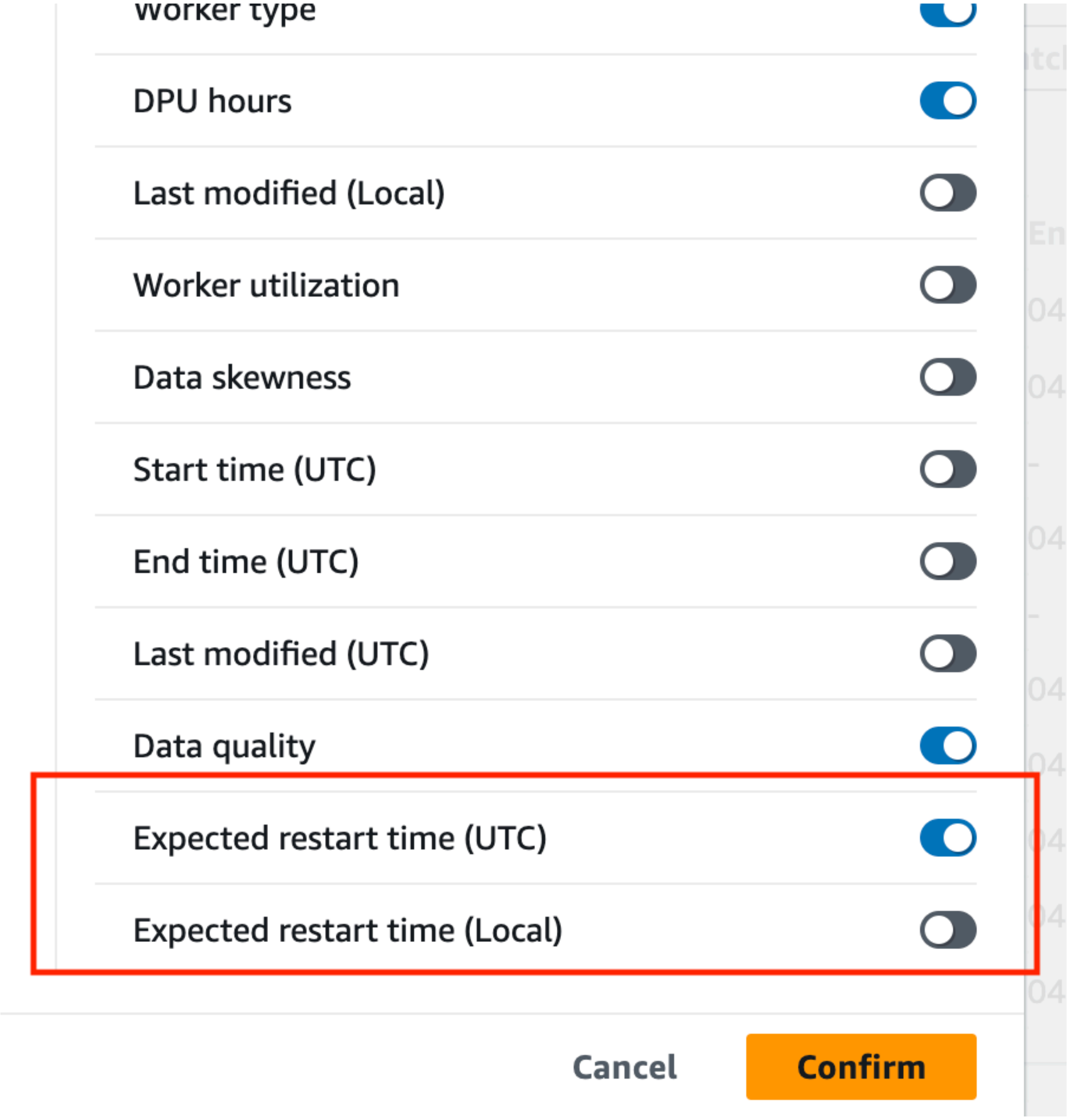
To see the expected next restart time of streaming jobs, show the column on the Job runs table on the **Monitoring** page.

1. Click the Gear icon in the top right of the table.



	Job name	Run status	Data quality status	Type	Start time (Local)	End time (Local)	Run time
○	auto-dq-test	Succeeded		Glue ETL	04/18/2024 10:23:26	04/18/2024 10:25:02	1 minute

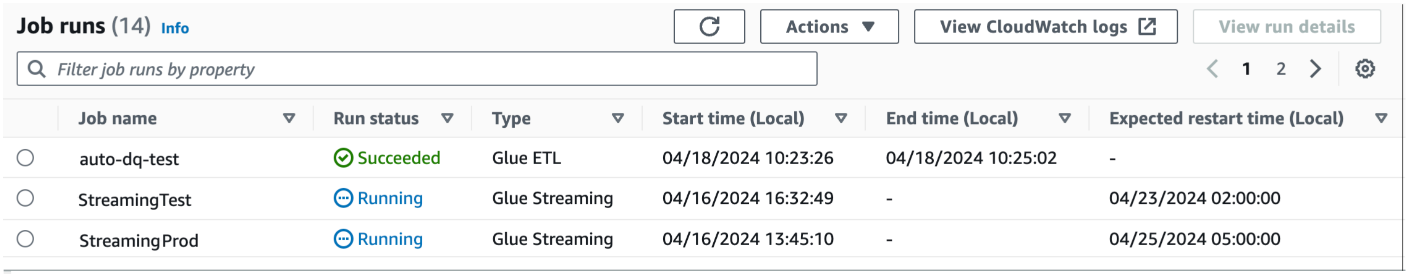
2. Scroll down, and turn on the **Expected restart time** column. Both UTC and Local time options are available.



Worker type	☑
DPU hours	☑
Last modified (Local)	☐
Worker utilization	☐
Data skewness	☐
Start time (UTC)	☐
End time (UTC)	☐
Last modified (UTC)	☐
Data quality	☑
Expected restart time (UTC)	☑
Expected restart time (Local)	☐

Cancel
Confirm

3. You can then view the columns in the table.



Job runs (14) Info

Actions
View CloudWatch logs
View run details
Filter job runs by property
1
2

	Job name	Run status	Type	Start time (Local)	End time (Local)	Expected restart time (Local)
○	auto-dq-test	✓ Succeeded	Glue ETL	04/18/2024 10:23:26	04/18/2024 10:25:02	-
○	StreamingTest	Running	Glue Streaming	04/16/2024 16:32:49	-	04/23/2024 02:00:00
○	StreamingProd	Running	Glue Streaming	04/16/2024 13:45:10	-	04/25/2024 05:00:00

The original job will have an "EXPIRED" status and the new job instance will have a "RUNNING" status. The new job run that was restarted will have a job run ID as a concatenation of initial job run ID plus the prefix "restart_" representing the restart count. For example, if the initial job run ID is `jr_1234`, then the restarted job run will have the ID `jr1234_restart_1` for the first restart. The second restart will be `jr1234_restart_2` for the second restart and so on.

Your retry attempt will not be impacted because of the restarts. If a run fails and a new run is started due to an automatic retry, the counter of restart will start from 1 again . For example, if a run fails at `jr_1234_attempt_3_restart_5`, then an automatic retry will start new run with ID: `jr_id1_attempt_4` and when this attempt is restarted after 7 days, the new run ID will be `jr_id1_attempt_4_restart_1`.

Data loss handling

During maintenance restarts, AWS Glue Streaming follows a process that ensures data integrity and consistency between the previous job run and the restarted job run. Note that AWS Glue does not guarantee data integrity and consistency between job restarts and we recommend architecture considerations to handle duplicated data within streaming jobs.

1. Detecting maintenance restart conditions: AWS Glue Streaming monitors conditions that indicate when a maintenance restart should be triggered, such as when a maintenance window is reached after 7 days or a hard restart is necessary after 14 days.
2. Invoking a graceful termination: When the maintenance restart conditions are met, AWS Glue Streaming initiates a graceful termination process for the currently running job. This process involves the following steps:
    a. Stopping the ingestion of new data: The streaming job stops consuming new data from the input sources (for example, Kafka topics, Kinesis streams, or files).
    b. Processing pending data: The job continues to process any data that is already present in its internal buffers or queues.
    c. Committing offsets and checkpoints: The job commits the latest offsets or checkpoints to external systems (for example, Kafka, Kinesis, or Amazon S3) to ensure that the restarted job can pick up from where the previous job left off.
3. Restarting the job: After the graceful termination process is complete, AWS Glue Streaming restarts the job using the preserved state and checkpoints. The restarted job picks up processing from the last committed offset or checkpoint, ensuring that no data is lost or duplicated.

4. Resuming data processing: The restarted job resumes data processing from the point where the previous job left off. It continues ingesting new data from the input sources, starting from the last committed offset or checkpoint, and processes the data according to the defined ETL logic.

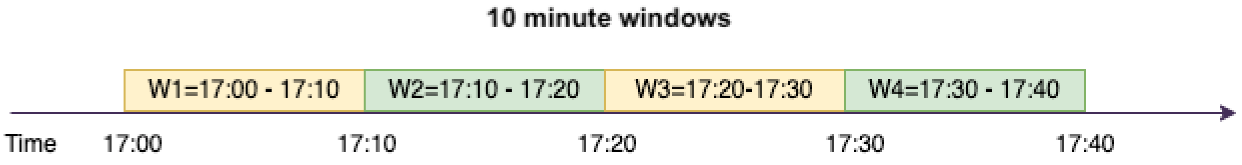
Advanced AWS Glue streaming concepts

In contemporary data-driven applications, the significance of data diminishes over time and its value transitions from being predictive to reactive. As a result, customers want to process data in real-time for making faster decisions. When dealing with real-time data feeds, such as from IoT sensors, the data may arrive unordered or experience delays in processing due to network latency and other source-related failures during ingestion. As part of the AWS Glue platform, AWS Glue Streaming builds on these capabilities to provide scalable, serverless streaming ETL, powered by Apache Spark structured streaming, empowering users with real-time data processing.

In this topic, we will explore advanced streaming concepts and capabilities of AWS Glue Streaming.

Time considerations when processing streams

There are four notions of time when processing streams:



- **Event-time** – The time at which the event occurred. In most cases, this field is embedded into the event-data itself, at the source.
- **Event-time-window** – The time frame between two event-times. As shown in the above diagram, **W1** is an event-time-window from 17:00 to 17:10. Each event-time-window is a grouping of multiple events.
- **Trigger-time** – The trigger time controls how often the processing of data and updating of results occurs. This is the time when the micro-batch processing started.
- **Ingestion-time** – The time when the stream-data was ingested into the streaming service. If event-time is not embedded into the event itself, this time can be used for windowing in some cases.

Windowing

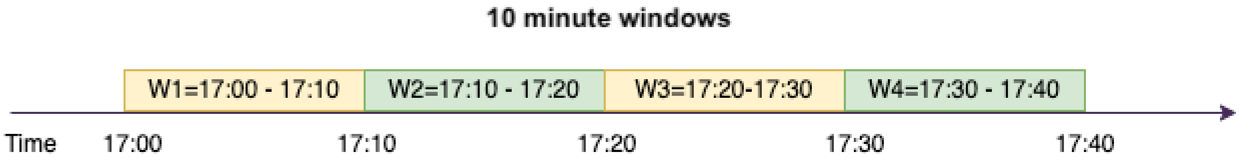
Windowing is a technique where you group and aggregate multiple events by event-time-window. We will explore the benefits of windowing and when you would use it in the following examples.

Depending on the business use case, there are three types of time windows supported by spark.

- Tumbling window – a series of non-overlapping fixed size event-time-windows over which you aggregate.
- Sliding window – similar to the tumbling windows from the point of being “fixed-sized”, but windows can overlap or slide as long as the duration of slide is smaller than the duration of window itself.
- Session window – starts with an input data event and continues to expands itself as long as it receives input within a gap or duration of inactivity. A session window can have a static or dynamic size of the window length, depending on the inputs.

Tumbling window

Tumbling window is a series of non-overlapping fixed size event-time-windows over which you aggregate. Lets understand this with a real world example.



Company ABC Auto wants to do a marketing campaign for a new brand of sports car. They want to pick a city where they have biggest sports car fans. To achieve this goal, they showcase a short 15 second advertisement introducing the car on their website. All the “clicks“ and the corresponding ”city“ are recorded and streamed to Amazon Kinesis Data Streams. We want to count the number of clicks in a 10 minute window and group it by city to see which city has the highest demand. The following is the output of the aggregation.

window_start_time	window_end_time	city	total_clicks
2023-07-10 17:00:00	2023-07-10 17:10:00	Dallas	75

window_start_time	window_end_time	city	total_clicks
2023-07-10 17:00:00	2023-07-10 17:10:00	Chicago	10
2023-07-10 17:20:00	2023-07-10 17:30:00	Dallas	20
2023-07-10 17:20:00	2023-07-10 17:30:00	Chicago	50

As explained above, these event-time-windows are different from trigger-time intervals. For example, even if your trigger time is every minute, the output results will only show 10 minute non-overlapping aggregation windows. For optimization, its better to have the trigger interval aligned with the event-time-window.

In the table above, Dallas saw 75 clicks in the 17:00-17:10 window, while Chicago had 10 clicks. Also, there is no data for the 17:10 - 17:20 window for any city, so this window is omitted.

Now you can run further analysis on this data in the downstream analytics application to determine the most exclusive city to run the marketing campaign.

Using tumbling windows in AWS Glue

1. Create a Amazon Kinesis Data Streams DataFrame and read from it. Example:

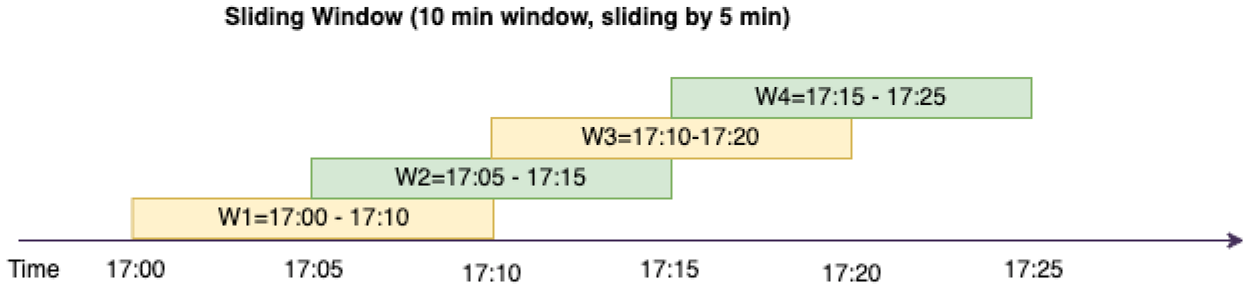
```
parsed_df = kinesis_raw_df \
                .selectExpr('CAST(data AS STRING)') \
                .select(from_json("data", ticker_schema).alias("data")) \
                .select('data.event_time','data.ticker','data.trade','data.volume',
 'data.price')
```

2. Process data in a tumbling window. In the example below, data is grouped based on the input field “event_time” in 10 minute tumbling windows and writing the output to an Amazon S3 data lake.

```
grouped_df = parsed_df \
                .groupBy(window("event_time", "10 minutes"), "city") \
                .agg(sum("clicks").alias("total_clicks"))

summary_df = grouped_df \
                .withColumn("window_start_time", col("window.start")) \
                .withColumn("window_end_time", col("window.end")) \
                .withColumn("year", year("window_start_time")) \
                .withColumn("month", month("window_start_time")) \
                .withColumn("day", dayofmonth("window_start_time")) \
                .withColumn("hour", hour("window_start_time")) \
                .withColumn("minute", minute("window_start_time")) \
                .drop("window")

write_result = summary_df \
                .writeStream \
                .format("parquet") \
                .trigger(processingTime="10 seconds") \
                .option("checkpointLocation", "s3a://bucket-stock-stream/stock-
stream-catalog-job/checkpoint/") \
                .option("path", "s3a://bucket-stock-stream/stock-stream-catalog-
job/summary_output/") \
                .partitionBy("year", "month", "day") \
                .start()
```

Sliding window

Sliding windows are similar to the tumbling windows from the point of being “fixed-sized”, but windows can overlap or slide as long as the duration of slide is smaller than the duration of window itself. Due to the nature of sliding, an input can be bound to the multiple windows.



To better understand, lets consider the example of a bank that want to detect potential credit card fraud. A streaming application could monitor a continuous stream of credit card transactions. These transactions could be aggregated into windows of 10 minutes duration and every 5 minutes, the window would slide forward, eliminating the oldest 5 minutes of data and adding the latest 5 minutes of new data. Within each window, the transactions could be grouped by country checking for suspicious patterns, such as a transaction in the US immediately followed by another in

Australia. For simplicity, lets us categorize such transactions as fraud when the total transactions amount is greater than $100. If such a pattern is detected, it signals potential fraud and the card could be frozen.

The credit card processing system is sending a steam of transaction events to kinesis for each card-id along with the country. An AWS Glue job runs the analysis and produces the following aggregated output.

window_start_time	window_end_time	card_last_four	country	total_amount
2023-07-10 17:00:00	2023-07-10 17:10:00	6544	US	85
2023-07-10 17:00:00	2023-07-10 17:10:00	6544	Australia	10
2023-07-10 17:05:45	2023-07-10 17:15:45	6544	US	50
2023-07-10 17:10:45	2023-07-10 17:20:45	6544	US	50
2023-07-10 17:10:45	2023-07-10 17:20:45	6544	Australia	150

Based on the above aggregation, you can see the 10 minute window sliding every 5 minutes,summed by transaction amount. The anomaly is detected in the 17:10 - 17:20 window where there is an outlier, which is a transaction for $150 in Australia. AWS Glue can detect this anomaly and push an alarm-event with the offending key to an SNS topic using boto3. Further a Lambda function can subscribe to this topic and take action.

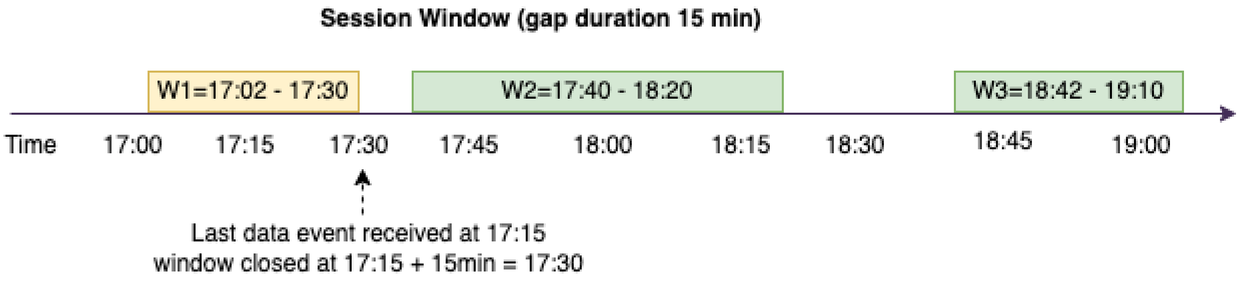
Process data in a sliding window

The `group-by` clause and the window function is used to implement the sliding window as shown below.

```
grouped_df = parsed_df \
            .groupBy(window(col("event_time"), "10 minute", "5 min"), "country",
 "card_last_four") \
            .agg(sum("tx_amount").alias("total_amount"))
```

Session window

Unlike the above two windows that have a fixed-size, session window can have a static or dynamic size of the window length, depending on the inputs. A session window starts with an input data event and continues to expands itself as long as it receives input within a gap or duration of inactivity.


Session Window (gap duration 15 min)
W1=17:02 - 17:30
W2=17:40 - 18:20
W3=18:42 - 19:10
Time
17:00
17:15
17:30
17:45
18:00
18:15
18:30
18:45
19:00
Last data event received at 17:15
window closed at 17:15 + 15min = 17:30

Lets take an example. Company ABC hotel wants to find out when is the busiest time in a week and provide better deals for their guests. As soon as a guest checks-in, a session window is started and spark maintains a state with aggregation for that event-time-window. Every time a guests checks in, an event is generated and sent to Amazon Kinesis Data Streams. The hotel makes a decision that if there is no check-ins for a period of 15 minutes, the event-time-window can be closed. The next event-time-window will start again when there is a new check-in. The output looks as follows.

window_start_time	window_end_time	city	total_checkins
2023-07-10 17:02:00	2023-07-10 17:30:00	Dallas	50
2023-07-10 17:02:00	2023-07-10 17:30:00	Chicago	25
2023-07-10 17:40:00	2023-07-10 18:20:00	Dallas	75
2023-07-10 18:50:45	2023-07-10 19:15:45	Dallas	20

The first check-in occurred at event_time=17:02. The aggregation event-time-window will start at 17:02. This aggregation will continue as long as we receive events within 15 minute duration. In the above example, the last event we received was at 17:15 and then for the next 15 minutes there were no events. As a result, Spark closed that event-time-window at 17:15+15min = 17:30 and set it as 17:02 - 17:30. It started a new event-time-window at 17:47 when it received a new check-in data event.

Process data in a session window

The `group-by` clause and the window function is used to implement the sliding window.

```
grouped_df = parsed_df \
            .groupBy(session_window(col("event_time"), "10 minute"), "city") \
            .agg(count("check_in").alias("total_checkins"))
```

Output modes

Output mode is the mode in which the results from the unbounded table are written to the external sink. There are three modes available. In the following example you are counting occurrences of a word as lines of data are being streamed and processed in each micro batch.

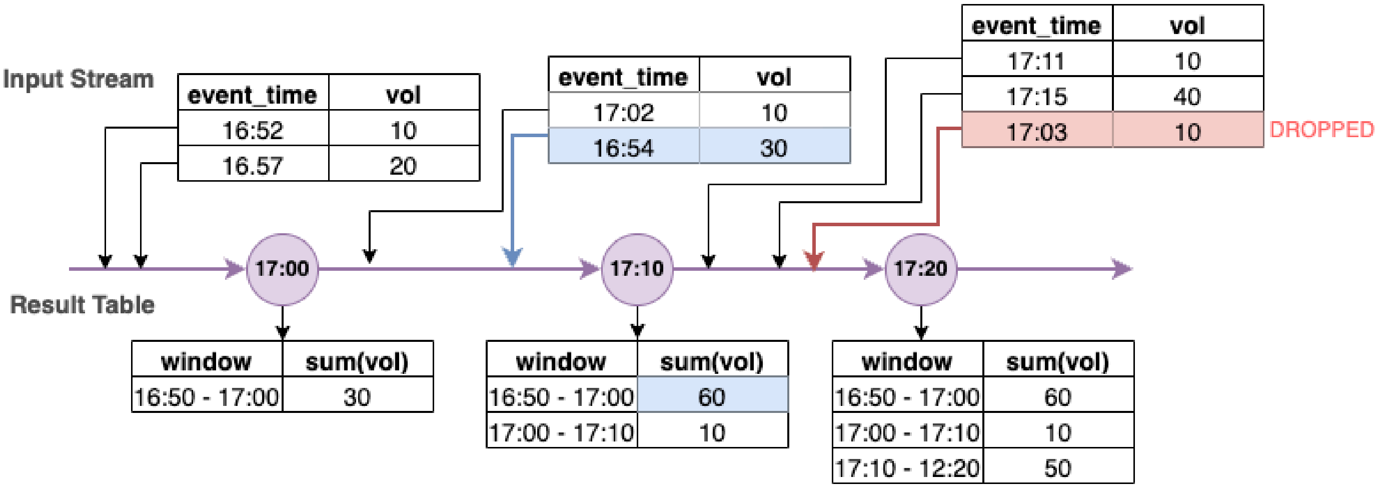
- **Complete mode** – The whole result table will be written to the sink after every micro batch processing even though the word count was not updated in the current event-time-window.
- **Append mode** – This is the default mode, where only the new words and or rows added to the result table since the last trigger will be written to the sink. This mode is good for stateless streaming for queries like map, flatMap, filter, etc.
- **Update mode** – Only the words and or rows in the Result Table that were updated or added since the last trigger will be written to the sink.

> **Note**
>
> Output mode = "update" is not supported for session windows.

Handling late data and watermarks

When working with real-time data there could be delays in the arrival of data due to network latency and upstream failures and we need a mechanism to perform the aggregation again on the missed event-time-window. However, to do this, state needs to be maintained. At the same time, the older data needs to be cleaned up to limit the size of the state. Spark version 2.1 added support for a feature called watermarking which maintains state and allows the user to specify the threshold for late data.

With reference to our stock ticker example above, lets consider the allowed threshold for the late data as no more than 10 minutes. To keep it simple we will assume tumbling window, ticker as AMZ, trade as BUY.


Input Stream
event_time | vol
16:52 | 10
16.57 | 20
event_time | vol
17:02 | 10
16:54 | 30
event_time | vol
17:11 | 10
17:15 | 40
17:03 | 10
DROPPED
17:00
17:10
17:20
Result Table
window | sum(vol)
16:50 - 17:00 | 30
window | sum(vol)
16:50 - 17:00 | 60
17:00 - 17:10 | 10
window | sum(vol)
16:50 - 17:00 | 60
17:00 - 17:10 | 10
17:10 - 12:20 | 50

In the above diagram, we are calculating the total volume over a tumbling 10 minute window. We have the trigger at 17:00, 17:10 and 17:20. Above the timeline arrow, we have the input data stream and below is the unbounded results table.

In the first 10 minute tumbling window we aggregated based on event_time and the total_volume was calculated as 30. In the second event-time-window, spark got the first data event with event_time=17:02. Since this is the max event_time seen thus far by spark, the watermark threshold is set 10 minutes back (that is, watermark_event_time=16:52). Any data event with an event_time after 16:52 will be considered for time bound aggregation and any data event before that will be dropped. This allows spark to maintain an intermediate state for additional 10 minutes to accommodate late data. Around wall clock time 17:08 Spark received an event with an event_time=16:54 which was within threshold. Hence spark recalculated the “16:50 - 17:00“ event-time-window and the total volume was updated from 30 to 60.

However, at the trigger time 17:20, when spark received event with event_time=17:15 it set the watermark_event_time=17:05. Hence the late data event with event_time=17:03 was considered “too late” and ignored.

```
Watermark Boundary = Max(Event Time) - Watermark Threshold
```

Using watermarks in AWS Glue

Spark will not emit or write the data to the external sink until the watermark boundary is passed. To implement a watermark in AWS Glue, see the example below.

```
grouped_df = parsed_df \
                .withWatermark("event_time", "10 minutes") \
                .groupBy(window("event_time", "5 minutes"), "ticker") \
                .agg(sum("volume").alias("total_volume"))
```

About monitoring AWS Glue streaming jobs

Monitoring your streaming job is a critical part of building your ETL pipeline. Apart from using the Spark UI, you can also use Amazon CloudWatch to monitor the metrics. Below is a list of streaming metrics emitted by the AWS Glue framework. For a complete list of all the AWS Glue metrics, see Monitoring AWS Glue using Amazon CloudWatch metrics.

AWS Glue uses a structured streaming framework to process the input events. You can either use the Spark API directly in your code or leverage the `ForEachBatch` provided by `GlueContext`, which publishes these metrics. To understand these metrics, we need to first understand `windowSize`.

windowSize: `windowSize` is the micro-batch interval that you provide. If you specify a window size of 60 seconds, the AWS Glue streaming job will wait for 60 seconds (or more if the previous batch hasn’t completed by then) before it will read data in a batch from the streaming source and apply the transformations provided in `ForEachBatch`. This is also referred to as the trigger interval.

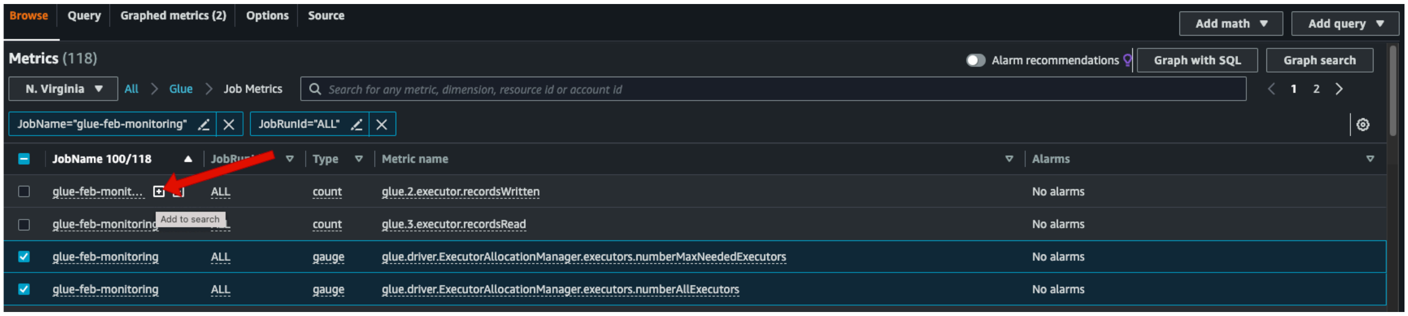
Lets review the metrics in greater detail to understand the health and performance characteristics.

> **Note**
>
> The metrics are emitted every 30 seconds. If your `windowSize` is less than 30 seconds then the reported metrics are an aggregation. For example say your `windowSize` is 10 seconds and you are steadily processing 20 records per micro-batch. In this scenario, the emitted metric value for numRecords would be 60.
> A metric is not emitted if there is no data available for it. Also, in case of the consumer lag metric, you have to enable the feature to get metrics for it.

Visualizing AWS Glue streaming metrics

To plot visual metrics:

1. Go to **Metrics** in the Amazon CloudWatch console and then choose the **Browse** tab. Then choose **Glue** under "Custom namespaces".


Browse
Query
Graphed metrics (2)
Options
Source
Add math
Add query
Alarm recommendations
Graph with SQL
Graph search
Add to search

Metrics (118)

N. Virginia | All > Glue > Job Metrics

Search for any metric, dimension, resource id or account id

JobName="glue-feb-monitoring"

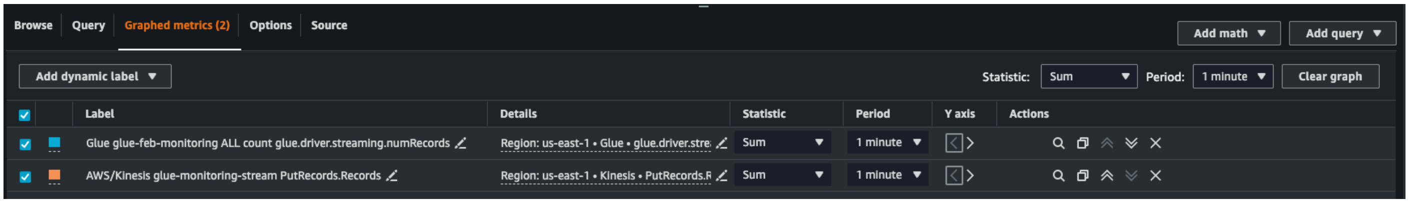
JobRunId="ALL"

	JobName 100/118	JobRu...	Type	Metric name	Alarms
☐	glue-feb-monit...	ALL	count	glue.2.executor.recordsWritten	No alarms
☐	glue-feb-monitoring	...	count	glue.3.executor.recordsRead	No alarms
☑	glue-feb-monitoring	ALL	gauge	glue.driver.ExecutorAllocationManager.executors.numberMaxNeededExecutors	No alarms
☑	glue-feb-monitoring	ALL	gauge	glue.driver.ExecutorAllocationManager.executors.numberAllExecutors	No alarms

2. Choose **Job Metrics** to show you the metrics for all your jobs.

3. Filter the metrics based on your JobName=glue-feb-monitoring and then JobRunId=ALL. You can click on the "+" sign as shown in the figure below to add it to the search filter.

4. Select the checkbox for the metrics that you are interested in. In the below figure we have selected `numberAllExecutors` and `numberMaxNeededExecutors`.


Browse
Query
Graphed metrics (2)
Options
Source
Add math
Add query
Add dynamic label
Clear graph

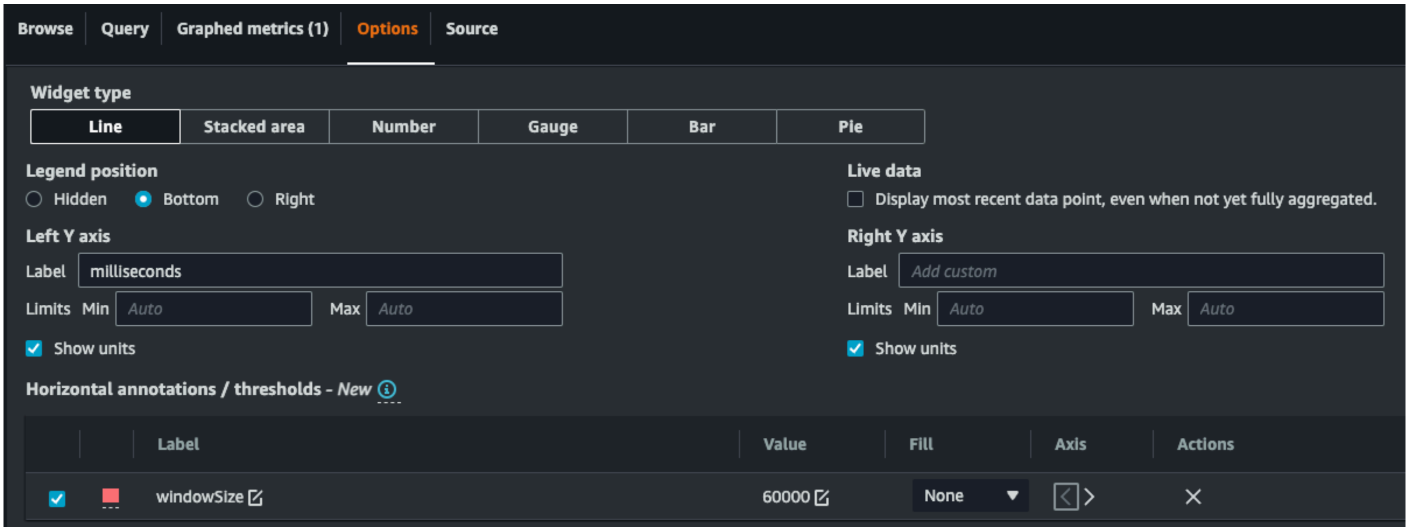
Statistic: Sum Period: 1 minute

☑		Label	Details	Statistic	Period	Y axis	Actions
☑	🟦	Glue glue-feb-monitoring ALL count glue.driver.streaming.numRecords	Region: us-east-1 • Glue • glue.driver.stre	Sum	1 minute	< >	
☑	🟧	AWS/Kinesis glue-monitoring-stream PutRecords.Records	Region: us-east-1 • Kinesis • PutRecords.R	Sum	1 minute	< >	

5. Once you have selected these metrics, you can go to the **Graphed metrics** tab and apply your statistics.

6. Since the metrics are emitted every min, you can apply the "average" over a minute for `batchProcessingTimeInMs` and `maxConsumerLagInMs`. For the `numRecords` you can apply the "sum" over every minute.

7. You can add a horizontal `windowSize` annotation to your graph using the **Options** tab.


Browse
Query
Graphed metrics (1)
Options
Source

Widget type

Line | Stacked area | Number | Gauge | Bar | Pie

Legend position

○ Hidden ● Bottom ○ Right

Live data

☐ Display most recent data point, even when not yet fully aggregated.

Left Y axis

Label: milliseconds

Limits Min: Auto Max: Auto

☑ Show units

Right Y axis

Label: Add custom

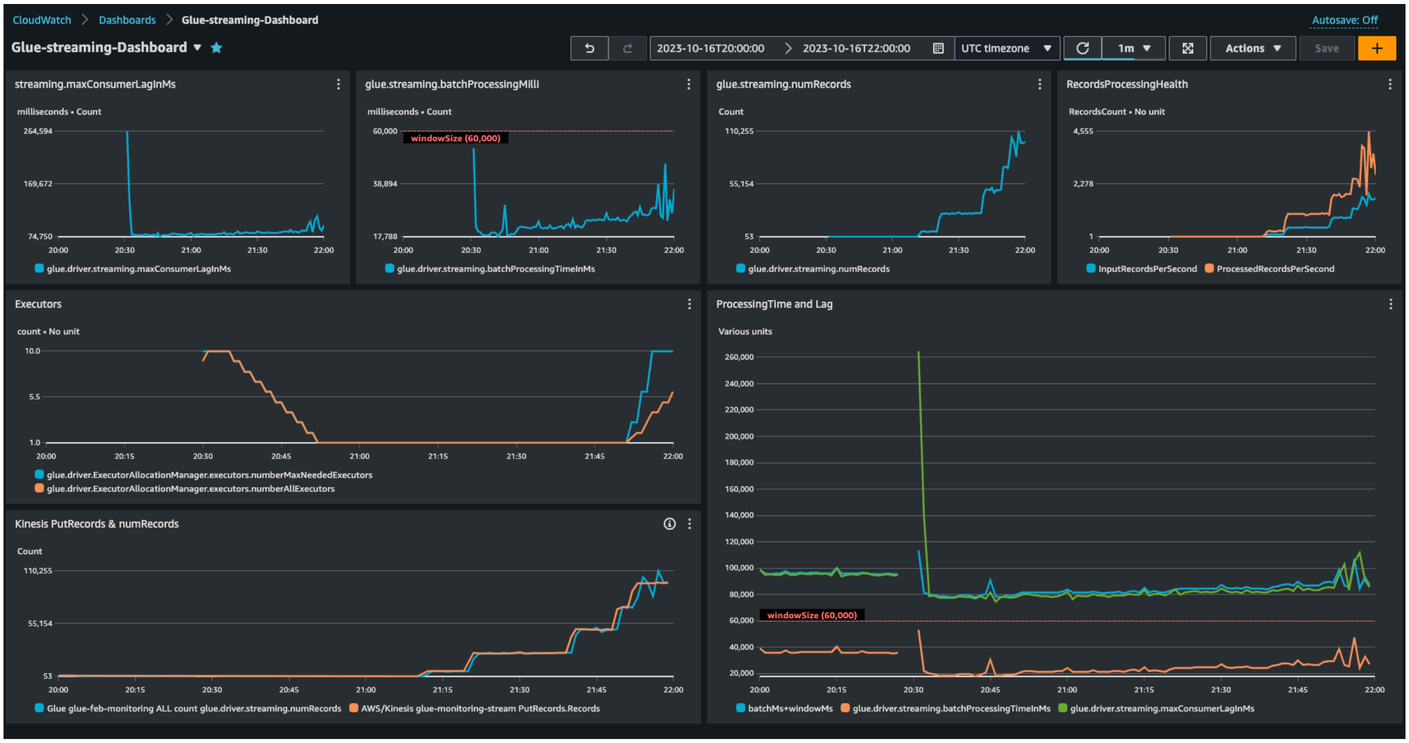
Limits Min: Auto Max: Auto

☑ Show units

Horizontal annotations / thresholds - *New*

		Label	Value	Fill	Axis	Actions
☑	🟥	windowSize	60000	None	< >	✕

8. Once you have your metrics selected, create a dashboard and add it. Here is a sample dashboard.

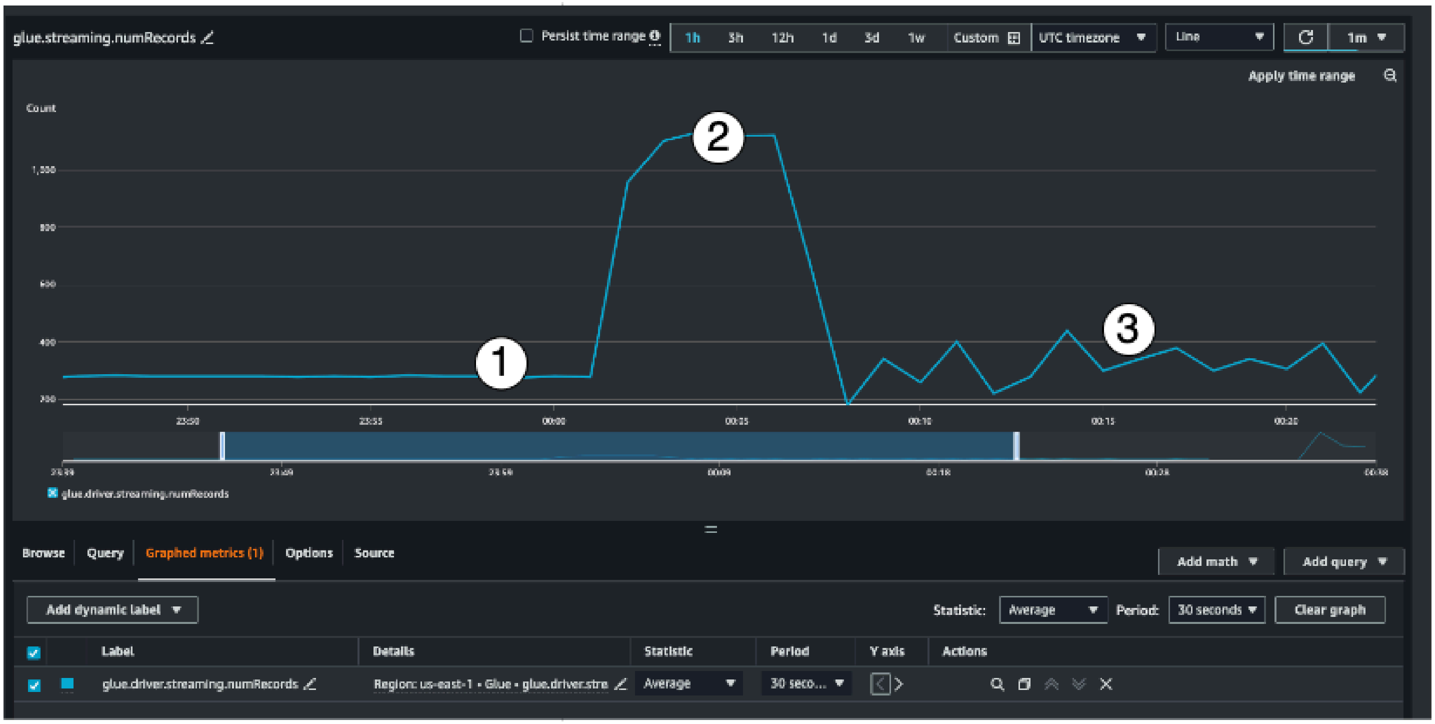

CloudWatch > Dashboards > Glue-streaming-Dashboard
Autosave: Off
Glue-streaming-Dashboard
2023-10-16T20:00:00 > 2023-10-16T22:00:00
UTC timezone
1m
Actions
Save
streaming.maxConsumerLagInMs
milliseconds • Count
264,594
169,672
74,750
20:00
20:30
21:00
21:30
22:00
glue.driver.streaming.maxConsumerLagInMs
glue.streaming.batchProcessingMilli
milliseconds • Count
60,000
windowSize (60,000)
38,894
17,788
glue.driver.streaming.batchProcessingTimeInMs
glue.streaming.numRecords
Count
110,255
55,154
53
glue.driver.streaming.numRecords
RecordsProcessingHealth
RecordsCount • No unit
4,555
2,278
1
InputRecordsPerSecond
ProcessedRecordsPerSecond
Executors
count • No unit
10.0
5.5
1.0
20:15
20:45
21:15
21:45
glue.driver.ExecutorAllocationManager.executors.numberMaxNeededExecutors
glue.driver.ExecutorAllocationManager.executors.numberAllExecutors
ProcessingTime and Lag
Various units
260,000
240,000
220,000
200,000
180,000
160,000
140,000
120,000
100,000
80,000
60,000
40,000
20,000
windowSize (60,000)
batchMs+windowMs
glue.driver.streaming.batchProcessingTimeInMs
glue.driver.streaming.maxConsumerLagInMs
Kinesis PutRecords & numRecords
Count
110,255
55,154
53
Glue glue-feb-monitoring ALL count glue.driver.streaming.numRecords
AWS/Kinesis glue-monitoring-stream PutRecords.Records

Using AWS Glue streaming metrics

This section describes each of the metrics and how they co-relate with each other.

Number of records (metric: streaming.numRecords)

This metric indicates how many records are being processed.


glue.streaming.numRecords
☐ Persist time range
1h
3h
12h
1d
3d
1w
Custom
UTC timezone
Line
1m
Apply time range
Count
1,000
800
600
400
200
23:50
23:55
00:00
00:05
00:10
00:15
00:20
1
2
3
glue.driver.streaming.numRecords
Browse
Query
Graphed metrics (1)
Options
Source
Add math
Add query
Add dynamic label
Clear graph

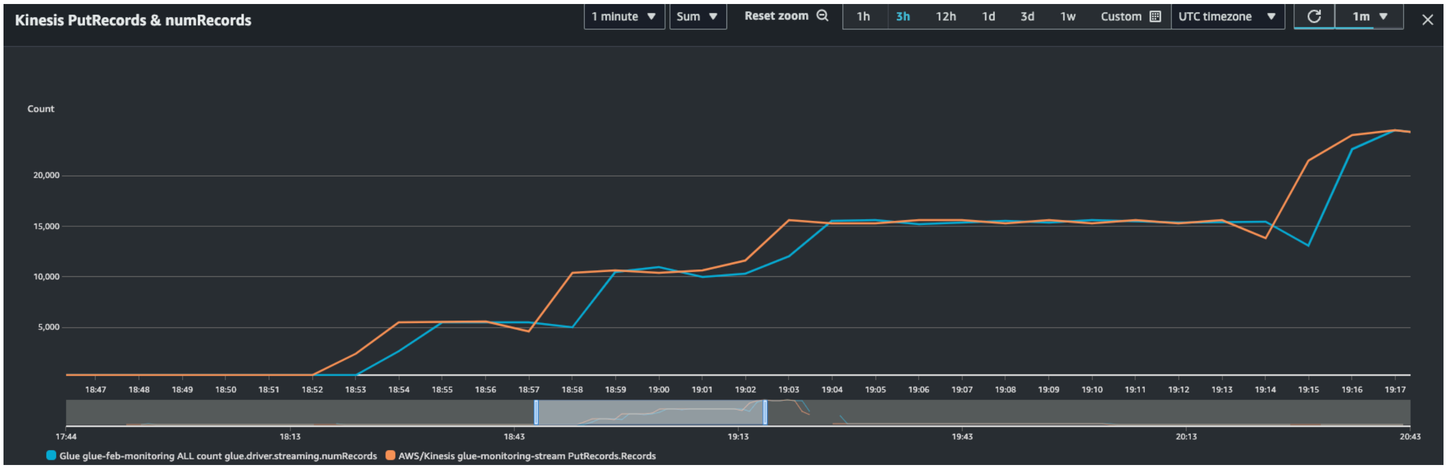
Statistic: Average

Period: 30 seconds

☑		Label	Details	Statistic	Period	Y axis	Actions
☑	🟦	glue.driver.streaming.numRecords	Region: us-east-1 • Glue • glue.driver.stre	Average	30 seco…		

This streaming metric provides visibility into the number of records you are processing in a window. Along with the number of records being processed, it will also help you understand the behavior of the input traffic.

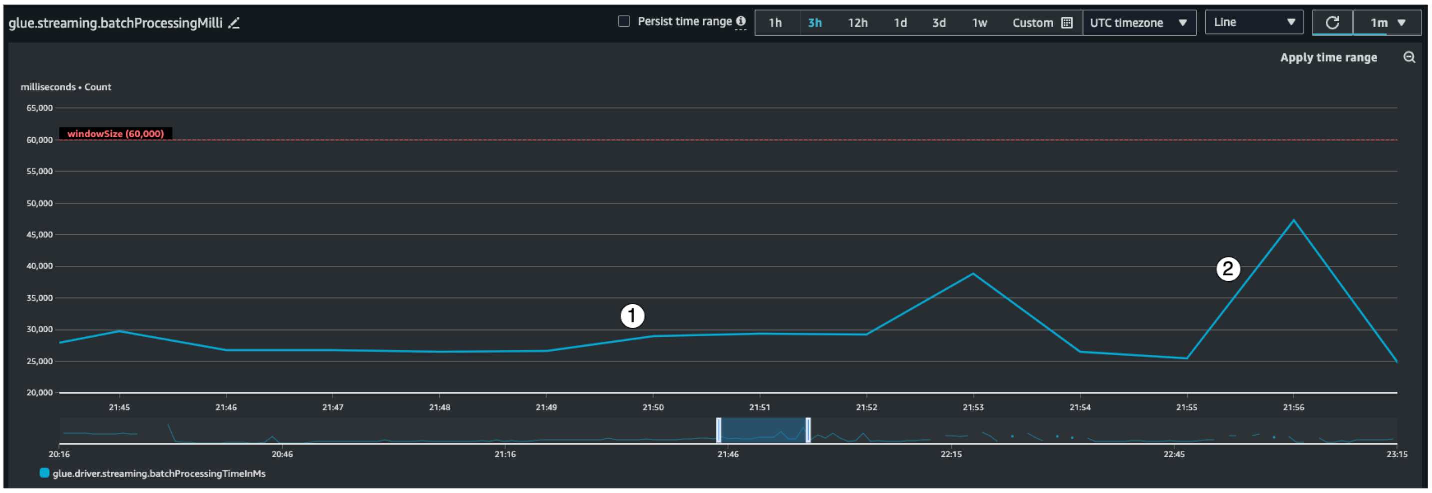
- Indicator #1 shows an example of stable traffic without any bursts. Typically this will be applications like IoT sensors that are collecting data at regular intervals and sending it to the streaming source.
- Indicator #2 shows an example of a sudden burst in traffic on an otherwise stable load. This could happen in a clickstream application when there is a marketing event like Black Friday and there is a burst in the number of clicks
- Indicator #3 shows an example of unpredictable traffic. Unpredictable traffic doesn't mean there is a problem. It is just the nature of the input data. Going back to the IoT sensor example, you can think of hundreds of sensors that are sending weather change events to the streaming source. As the weather change is not predictable, neither is the data. Understanding the traffic pattern is key to sizing your executors. If the input is very spiky, you may consider using autoscaling (more on that later).



You can combine this metric with the Kinesis PutRecords metric to make sure the number of events being ingested and the number of records being read are nearly the same. This is especially useful when you are trying to understand lag. As the ingestion rate increases, so do the `numRecords` read by AWS Glue.

Batch processing time (metric: streaming.batchProcessingTimeInMs)

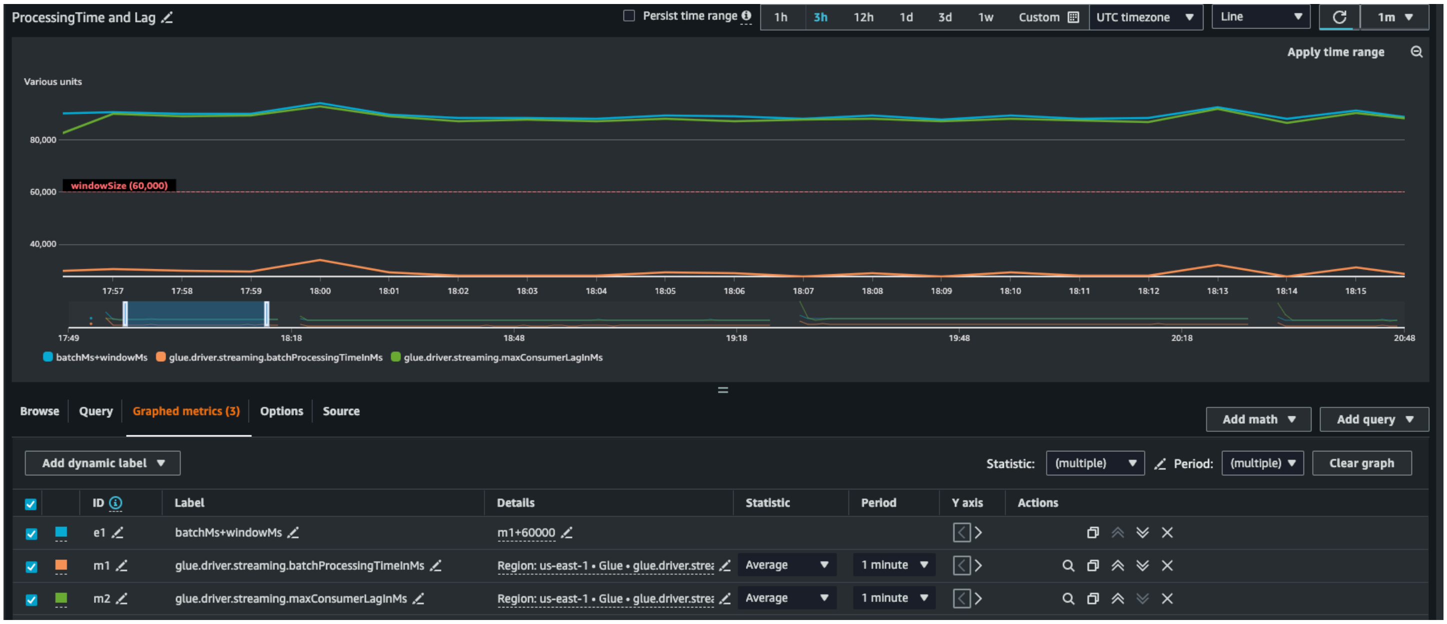
The batch processing time metric helps you determine if the cluster is underprovisioned or overprovisioned.



This metric indicates the number of milliseconds that it took to process each micro-batch of records. The main goal here is to monitor this time to make sure it less than the `windowSize` interval. It is okay if the `batchProcessingTimeInMs` goes over temporarily as long as it recovers in the following window interval. Indicator #1 shows a more or less stable time taken to process the job. However if the number of input records are increasing, the time it takes to process the job will increase as well as shown by indicator #2. If the `numRecords` is not going up, but the processing time is going up, then you would need to take a deeper look into the job processing on the executors. It is a good practice to set a threshold and alarm to make sure the `batchProcessingTimeInMs` doesn't stay over 120% for more than 10 minutes. For more information on setting alarms, see [Using Amazon CloudWatch alarms](#).

Consumer lag (metric: streaming.maxConsumerLagInMs)

The consumer lag metric helps you understand if there is a lag in processing events. If your lag is too high, then you could miss the processing SLA that your business depends on, even though you have a correct windowSize. You have to explicitly enable this metrics using the `emitConsumerLagMetrics` connection option. For more information, see [KinesisStreamingSourceOptions](#).

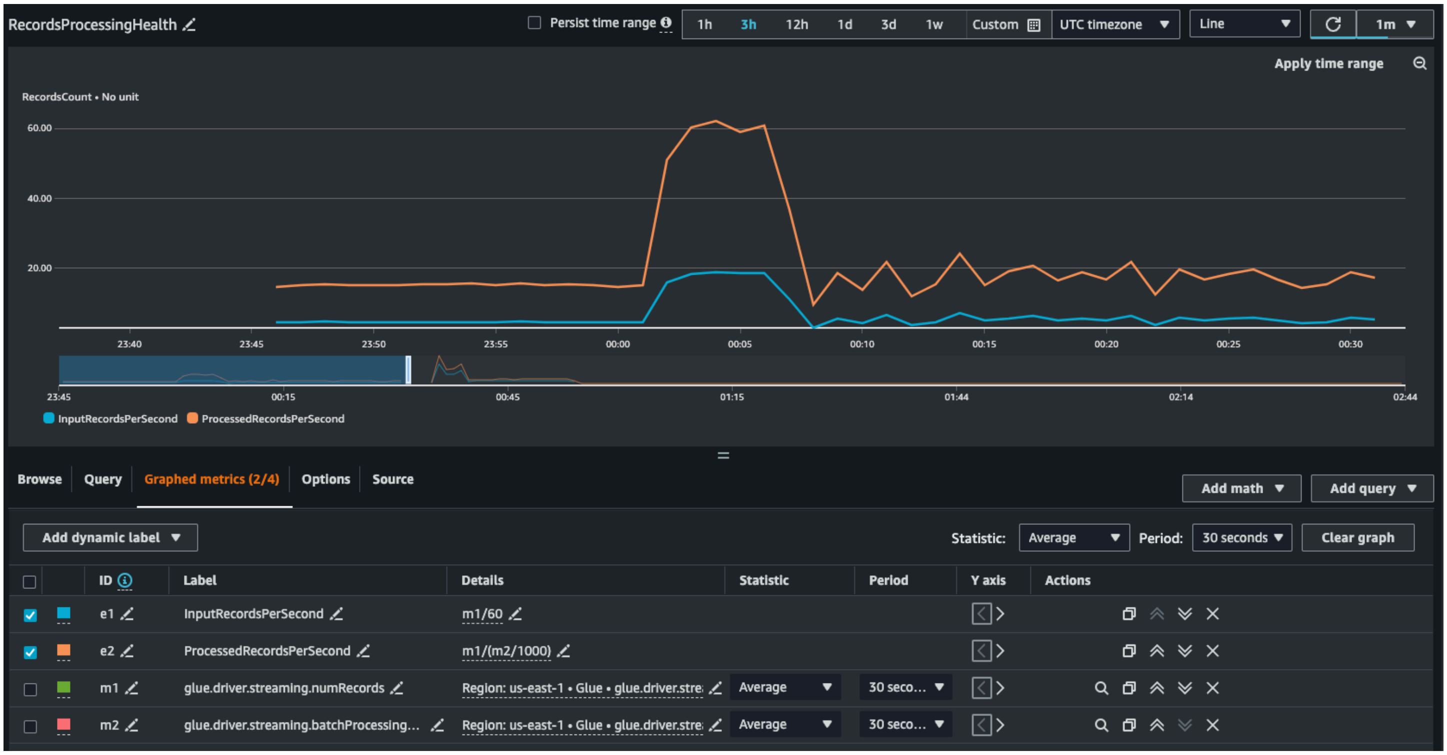

ProcessingTime and Lag
☐ Persist time range
1h
3h
12h
1d
3d
1w
Custom
UTC timezone
Line
1m
Apply time range
Various units
80,000
60,000
windowSize (60,000)
40,000
17:57
17:58
17:59
18:00
18:01
18:02
18:03
18:04
18:05
18:06
18:07
18:08
18:09
18:10
18:11
18:12
18:13
18:14
18:15
17:49
18:18
18:48
19:18
19:48
20:18
20:48
batchMs+windowMs
glue.driver.streaming.batchProcessingTimeInMs
glue.driver.streaming.maxConsumerLagInMs
Browse
Query
Graphed metrics (3)
Options
Source
Add math
Add query
Add dynamic label
Clear graph

Statistic: (multiple) Period: (multiple)

☑		ID	Label	Details	Statistic	Period	Y axis	Actions
☑	🟦	e1	batchMs+windowMs	m1+60000				
☑	🟧	m1	glue.driver.streaming.batchProcessingTimeInMs	Region: us-east-1 • Glue • glue.driver.strea	Average	1 minute		
☑	🟩	m2	glue.driver.streaming.maxConsumerLagInMs	Region: us-east-1 • Glue • glue.driver.strea	Average	1 minute		

Derived metrics

To gain deeper insights, you can create derived metrics to understand more about your streaming jobs in Amazon CloudWatch.


RecordsProcessingHealth
☐ Persist time range
1h
3h
12h
1d
3d
1w
Custom
UTC timezone
Line
1m
Apply time range
RecordsCount • No unit
60.00
40.00
20.00
23:40
23:45
23:50
23:55
00:00
00:05
00:10
00:15
00:20
00:25
00:30
23:45
00:15
00:45
01:15
01:44
02:14
02:44
InputRecordsPerSecond
ProcessedRecordsPerSecond
Browse
Query
Graphed metrics (2/4)
Options
Source
Add math
Add query
Add dynamic label
Statistic: Average
Period: 30 seconds
Clear graph

☐		ID	Label	Details	Statistic	Period	Y axis	Actions
☑	🟦	e1	InputRecordsPerSecond	m1/60				
☑	🟧	e2	ProcessedRecordsPerSecond	m1/(m2/1000)				
☐	🟩	m1	glue.driver.streaming.numRecords	Region: us-east-1 • Glue • glue.driver.stre	Average	30 seco...		
☐	🟥	m2	glue.driver.streaming.batchProcessing...	Region: us-east-1 • Glue • glue.driver.stre	Average	30 seco...		

You can build a graph with derived metrics to decide if you need to use more DPUs. While autoscaling helps you do this automatically, you can use derived metrics to determine if autoscaling is working effectively.

- **InputRecordsPerSecond** indicates the rate at which you are getting input records. It is derived as follows: number of input records (glue.driver.streaming.numRecords)/ WindowSize.
- **ProcessingRecordsPerSecond** indicates the rate at which your records are being processed. It is derived as follows: number of input records (glue.driver.streaming.numRecords)/ batchProcessingTimeInMs.

If the input rate is higher than the processing rate, then you may need to add more capacity to process your job or increase the parallelism.

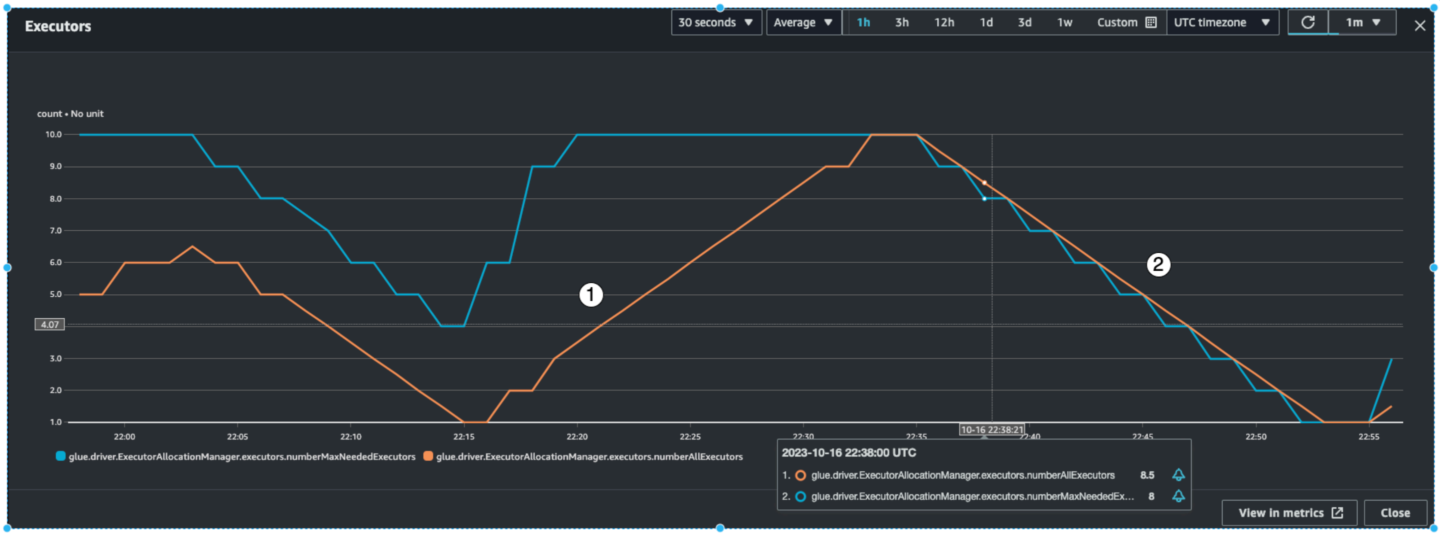
Autoscaling metrics

When your input traffic is spiky, then you should consider enabling autoscaling and specify the max workers. With that you get two additional metrics, `numberAllExecutors` and `numberMaxNeededExecutors`.

- **numberAllExecutors** is the number of actively running job executors

- **numberMaxNeededExecutors** is the number of maximum (actively running and pending) job executors needed to satisfy the current load.

These two metrics will help you understand if your autoscaling is working correctly.



AWS Glue will monitor the `batchProcessingTimeInMs` metric over a few micro-batches and do one of two things. It will scale-out the executors, if `batchProcessingTimeInMs` is closer to the `windowSize`, or scale-in the executors, if `batchProcessingTimeInMs` is comparatively lower than `windowSize`. Also, it will use an algorithm for step-scaling the executors.

- indicator #1 shows you how the active executors scaled up to catch up with the max needed executors so as to process the load.
- indicator #2 shows you how the active executers scaled in since the `batchProcessingTimeInMs` was low.

You can use these metrics to monitor current executor-level parallelism and adjust the number of max workers in your auto-scaling configuration accordingly.

How to get the best performance

Spark will try to create one task per shard, to read from, in the Amazon Kinesis stream. The data in each shard becomes a partition. It will then distribute these tasks across the executors/workers, depending of the number of cores on each worker (the number of cores per worker depends on the worker type you select `G.025X`, `G.1X`, etc). However it is non-deterministic how the tasks are distributed. All tasks are executed in parallel on their respective cores. If there are more shards than the number of available executor cores, the tasks are queued up.

You can use a combination of the above metrics and the number of shards, to provision your executors for a stable load with some room for bursts. It is recommend that you run a few iterations of your job in order to determine the approximate number of workers. For an unstable/ spiky workload you can do the same by setting up autoscaling and max workers.

Set the `windowSize` as per the SLA requirement of your business. For example, if your business requires that the processed data cannot be more than 120 seconds stale, then set your `windowSize` to at least 60 seconds such that your average consumer lag is less than 120 seconds (refer to the section on consumer lag above). From there depending on the `numRecords` and number of shards, plan for the capacity in DPUs making sure your `batchProcessingTimeInMs` is less than 70% of your `windowSize` most of the time.

> **Note**
>
> Hot shards can cause data skew which means that some shards/partitions are much bigger than the others. This may cause some tasks that are running in parallel to take longer time causing straggler tasks. As a result, the next batch can't start until all tasks from the previous one complete, this will impact the `batchProcessingTimeInMillis` and the max lag.

Zero-ETL integrations

Zero-ETL is a set of fully managed integrations by AWS that minimizes the need to build ETL data pipelines for common ingestion and replication use cases. It makes data available in Amazon SageMaker Lakehouse and Amazon Redshift from multiple operational, transactional, and application sources. With zero-ETL integration, you have fresher data for analytics, AI/ML, and reporting. You get more accurate and timely insights for use cases like business dashboards, optimized gaming experience, data quality monitoring, and customer behavior analysis. You can make data-driven predictions with more confidence, improve customer experiences, and promote data-driven insights across the business.

Amazon Redshift is a fast, fully managed, petabyte-scale data warehouse service that makes it simple and cost-effective to efficiently analyze all your data using your existing business intelligence tools.

Amazon SageMaker Lakehouse unifies all your data across Amazon Simple Storage Service (S3) data lakes and Amazon Redshift data warehouses, helping you build powerful analytics and AI/ML applications on a single copy of data. SageMaker Lakehouse gives you the flexibility to access and query your data in-place with all Apache Iceberg compatible tools and engines. With SageMaker Lakehouse, you also have the flexibility to access and query your data in-place with Apache Iceberg compatible tools and engines. Additionally, you can secure your data with integrated, fine-grained access controls, that are enforced across all your data in all analytic tools and engines. Define permissions once and confidently share data across your organization.

Zero-ETL capabilities in AWS Glue

Zero-ETL integrations in AWS Glue simplify data ingestion and replication from AWS data services and third-party applications to AWS destinations.

AWS services supported by zero-ETL sources in AWS Glue include:

- Amazon DynamoDB

Third-party application supported by zero-ETL include:

- Facebook Ads
- Instagram Ads

- Salesforce
- Salesforce Marketing Cloud Account Engagement
- SAP OData
- ServiceNow
- Zendesk
- Zoho CRM

AWS services supported by zero-ETL targets in AWS Glue include:

- Amazon Redshift
- Amazon SageMaker Lakehouse

> **Note**
>
> When creating a zero-ETL integration with an Amazon DynamoDB source in AWS Glue, the target is supported by Amazon SageMaker Lakehouse.

Prerequisites for setting up a zero-ETL integration

Setting up an integration between the source and target require some prerequisites such as configuring IAM roles which AWS Glue uses to access data from the source and write to the target, and the use of KMS keys to encrypt the data in intermediate or the target location.

Topics

- Setting up source resources
- Setting up target resources
- Creating an Amazon Redshift data warehouse
- Setting up a VPC for your zero-ETL integration
- Setting up a zero-ETL cross-account integration

Setting up source resources

Perform the following set up tasks as required for your source.

Setting up the source role

This section describe how you pass a source role to allow the zero-ETL integration to access your connection. This is also applicable only for SaaS sources.

> **Note**
>
> To restrict access to only a few connections, you can first create the connection to obtain the connection ARN. See Configuring a source for a zero-ETL integration.

Create a role which has permissions for the integration to access the connection:

JSON

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "GlueConnections",
            "Effect": "Allow",
            "Action": [
                "glue:GetConnections",
                "glue:GetConnection"
            ],
            "Resource": [
       "arn:aws:glue:*:111122223333:catalog",
       "arn:aws:glue:us-east-1:111122223333:connection/*"
            ]
        },
        {
            "Sid": "GlueActionBasedPermissions",
            "Effect": "Allow",
            "Action": [
                "glue:ListEntities",
                "glue:RefreshOAuth2Tokens"
            ],
            "Resource": [
                "*"
            ]
        }
    ]
}
```

Trust policy:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {
        "Service": [
          "glue.amazonaws.com"
        ]
      },
      "Action": "sts:AssumeRole"
    }
  ]
}
```

Setting up target resources

Perform the following set up tasks as required for AWS Glue Data Catalog or Amazon Redshift data warehouse integration target.

For integrations with an AWS Glue database target:

- Setting up an AWS Glue database
- Providing a catalog Resource Based Access (RBAC) policy
- Creating a target IAM role

For integrations with an Amazon Redshift datawarehouse target:

- https://docs.aws.amazon.com/glue/latest/dg/zero-etl-prerequisites.html#zero-etl-setup-target-redshift-data-warehouse

Setting up an AWS Glue database

For integrations that use an AWS Glue database:

To set up a target database in the AWS Glue Data Catalog with an Amazon S3 location:

1. In the AWS Glue console home page, select **Database** under Data Catalog.
2. Choose **Add database** in the top right corner. If you have already created a database, make sure that the location with Amazon S3 URI is set for the database.
3. Enter a name and **Location** (Amazon S3 URI). Note that the location is required for the zero-ETL integration. Click **Create database** when done.

> **Note**
>
> The Amazon S3 bucket must be in the same region as the AWS Glue database.

For information on creating a new database in AWS Glue, see Getting started with the AWS Glue Data Catalog.

You can also use the `create-database` CLI to create the database in AWS Glue. Note that the `LocationUri` in `--database-input` is required.

Optimizing Iceberg tables

Once a table is created by AWS Glue in the target database, you can enable the compaction to speed up queries in Amazon Athena. For information on setting up the resources (IAM Role) for compaction, see Table optimization prerequisites.

For more information on setting up compaction on the AWS Glue table created by the integration, see Optimizing Iceberg tables.

Providing a catalog Resource Based Access (RBAC) policy

For integrations that use an AWS Glue database, add the following permissions to the catalog RBAC Policy to allow for integrations between source and target.

> **Note**
>
> For cross-account integrations, both Alice (user creating the integration) role policy and catalog resource policy need to allow `glue:CreateInboundIntegration`

on the resource. For same-account, either a resource policy or role policy allowing `glue:CreateInboundIntegration` on the resource is sufficient. Both scenarios do still need to allow `glue.amazonaws.com` to `glue:AuthorizeInboundIntegration`.

You can access the **Catalog settings** under **Data Catalog**. Then provide the following permissions and fill in the missing information.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Principal": {
        "AWS": [
         "arn:aws:iam::123456789012:user/Alice"
        ]
      },
      "Effect": "Allow",
      "Action": [
        "glue:CreateInboundIntegration"
      ],
      "Resource": [
       "arn:aws:glue:us-east-1:111122223333:catalog",
    "arn:aws:glue:us-east-1:111122223333:database/DatabaseName"
      ],
      "Condition": {
        "StringLike": {
        "aws:SourceArn": "arn:aws:dynamodb:us-east-1:444455556666:table/<table-
name>"
        }
      }
    },
    {
      "Principal": {
        "Service": [
          "glue.amazonaws.com"
        ]
      },
      "Effect": "Allow",
      "Action": [
        "glue:AuthorizeInboundIntegration"
      ],
      "Resource": [
       "arn:aws:glue:us-east-1:111122223333:catalog",
    "arn:aws:glue:us-east-1:111122223333:database/DatabaseName"
      ],
      "Condition": {
        "StringEquals": {
        "aws:SourceArn": "arn:aws:dynamodb:us-east-1:444455556666:table/<table-name>"
        }
      }
    }
  ]
}
```

Creating a target IAM role

Create a target IAM role with the following permissions and trust relationships:

JSON

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Action": "s3:ListBucket",
            "Resource": "arn:aws:s3:::amzn-s3-demo-bucket",
            "Effect": "Allow"
        },
        {
            "Action": [
                "s3:GetObject",
                "s3:PutObject",
                "s3:DeleteObject"
            ],
            "Resource": "arn:aws:s3:::amzn-s3-demo-bucket/prefix/*",
            "Effect": "Allow"
        },
        {
            "Action": [
                "glue:GetDatabase"
            ],
            "Resource": [
                "arn:aws:glue:us-east-1:111122223333:catalog",
                "arn:aws:glue:us-east-1:111122223333:database/DatabaseName"
            ],
            "Effect": "Allow"
        },
        {
            "Action": [
                "glue:CreateTable",
                "glue:GetTable",
                "glue:GetTables",
                "glue:DeleteTable",
                "glue:UpdateTable",
                "glue:GetTableVersion",
                "glue:GetTableVersions",
                "glue:GetResourcePolicy"
            ],
            "Resource": [
                "arn:aws:glue:us-east-1:111122223333:catalog",
                "arn:aws:glue:us-east-1:111122223333:database/DatabaseName",
                "arn:aws:glue:us-east-1:111122223333:table/DatabaseName/*"
            ],
            "Effect": "Allow"
        },
        {
            "Action": [
                "cloudwatch:PutMetricData"
            ],
            "Resource": "*",
            "Condition": {
                "StringEquals": {
                    "cloudwatch:namespace": "AWS/Glue/ZeroETL"
                }
            },
            "Effect": "Allow"
        },
        {
            "Action": [
                "logs:CreateLogGroup",
                "logs:CreateLogStream",
                "logs:PutLogEvents"
            ],
            "Resource": "*",
                "Effect": "Allow"
            }
        ]
}
```

Add the following trust policy to allow the AWS Glue service to assume the role:

JSON

```
{
    "Version": "2012-10-17",
    "Statement": [
      {
        "Effect": "Allow",
        "Principal": {
          "Service": "glue.amazonaws.com"
        },
        "Action": "sts:AssumeRole"
      }
    ]
}
```

Creating an Amazon Redshift data warehouse

When your zero-ETL integration target is an Amazon Redshift data warehouse, create the data warehouse if you don't already have one. To create an Amazon Redshift Serverless workgroup, see Creating a workgroup with a namespace. To create an Amazon Redshift cluster, see Creating a cluster.

The target Amazon Redshift workgroup or cluster must have the `enable_case_sensitive_identifier` parameter turned on for the integration to be successful. For more information on enabling case sensitivity, see Turn on case sensitivity for your data warehouse in the Amazon Redshift management guide.

After the Amazon Redshift workgroup or cluster setup is complete, you need to configure your data warehouse. See Getting started with zero-ETL integrations in the Amazon Redshift Management Guide for more information.

Setting up a VPC for your zero-ETL integration

To set up a VPC for your zero-ETL integration:

1. Go to **VPC** > Your VPCs and choose **Create VPC**.
    a. Select **VPC and more**.
    b. Set your VPC name.
    c. Set the IPv4 CIDR: 10.0.0.0/16.
    d. Set the number of AZ to 1.
    e. Set the number of public and private subnets to 1.
    f. Set **NAT gateways** to None.
    g. Set **VPC endpoints** to S3 Gateway.
    h. Enable DNS hostnames and DNS resolution.
2. Go to **Endpoints** and choose **Create Endpoint**.
3. Create endpoints for these services in the private subnet of your VPC (use the default security group):
    a. com.amazonaws.us-east-1.lambda
    b. com.amazonaws.us-east-1.glue
    c. com.amazonaws.us-east-1.sts

Create the AWS Glue connection:

1. Go to **AWS Glue** > **Data connections** and choose **Create connection**.
2. Select **Network**.
3. Select the VPC, Subnet (private), and default Security Group that you created.

Setting up the target role for the VPC

The target role must have these permissions (in addition to the other permissions required by Zero-ETl integrations):

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "CustomerVpc",
      "Effect": "Allow",
      "Action": [
        "ec2:CreateTags",
        "ec2:DeleteTags",
        "ec2:DescribeRouteTables",
        "ec2:DescribeVpcEndpoints",
        "ec2:DescribeSecurityGroups",
        "ec2:DescribeSubnets",
        "ec2:CreateNetworkInterface",
        "ec2:DeleteNetworkInterface",
        "glue:GetConnection"
      ],
      "Resource": [
        "*"
      ]
    }
  ]
}
```

Setting up the target leg resource properties

If you are using the CLI, set the target leg resource properties to the target AWS Glue database you created. Pass the target role ARN, as well as the AWS Glue connection name.

```
aws glue create-integration-resource-property \
--resource-arn arn:aws:glue:us-east-1:<account-id>:database/exampletarget \
--target-processing-properties '{"RoleArn" : "arn:aws:iam::<account-id>:role/example-role", "ConnectionName":"example-vpc-3"}' \
--endpoint-url https://example.amazonaws.com --region us-east-1
```

Possible client errors

The following are possible client errors for an integration configured with a VPC.

Error message	Action required
Provided role is not authorized to perform glue:GetConnection on connection. Add this permission to role policy, and then wait for the integration to recover.	Update role policy
Provided role is not authorized to perform ec2:DescribeSubnets. Add this permission to role policy, and then wait for the integration to recover.	Update role policy
Provided role is not authorized to perform ec2:DescribeSecurityGroups. Add this permission to role policy, and then wait for the integration to recover.	Update role policy
Provided role is not authorized to perform ec2:DescribeVpcEndpoints. Add this permission to role policy, and then wait for the integration to recover.	Update role policy
Provided role is not authorized to perform ec2:DescribeRouteTables. Add this permission to role policy, and then wait for the integration to recover.	Update role policy
Provided role is not authorized to perform ec2:CreateTags. Add this permission to role policy, and then wait for the integration to recover.	Update role policy
Provided role is not authorized to perform ec2:CreateNetworkInterface. Add this permission to role policy, and then wait for the integration to recover.	Update role policy

Error message	Action required
Provided connection subnet does not contain a valid S3 endpoint or NAT gateway. Update subnet, and then wait for the integration to recover.	Update VPC subnet endpoints
Connection subnet not found. Update connection subnet, and then wait for the integration to recover.	Update AWS Glue connection
Connection security group not found. Update connection security group, and then wait for the integration to recover.	Update AWS Glue connection
Can't connect to S3 through provided VPC connection. Update subnet configurations, and then wait for the integration to recover.	Update VPC subnet endpoints
Can't connect to Lambda through provided VPC connection. Update subnet configura tions, and then wait for the integration to recover.	Update VPC subnet endpoints

Setting up a zero-ETL cross-account integration

To set up a zero-ETL cross-account integration:

1. Configure a target Resource Policy as described in [Providing a catalog Resource Based Access (RBAC) policy](). Ensure that the source account role is explicitly allowed on the target resource.
2. Check that the source account role (the role used to create the integration) has the following:

   JSON

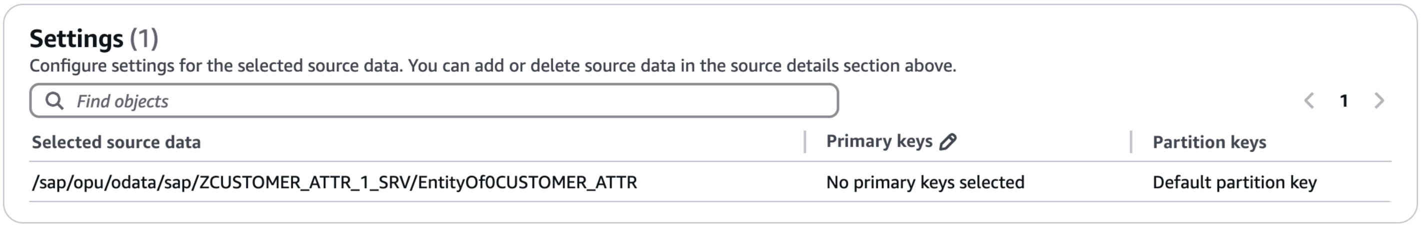
```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "Stmt123456789012",
            "Action": [
                "glue:CreateInboundIntegration"
            ],
            "Effect": "Allow",
            "Resource": [
             "arn:aws:glue:us-east-1:111122223333:catalog",
       "arn:aws:glue:us-east-1:111122223333:database/DatabaseName"
            ]
        }
    ]
}
```

3. Create the integration as described in [Creating an integration](#).

Configuring a source for a zero-ETL integration

Support for Special SAP Entities

For most of the SaaS entities, we determine the valid primary key sets during data processing, however there are some which require an extra step of providing the valid primary key set as input, specifically SAP entities that start with `EntityOf`. When an `EntityOf` entity is selected, you will be directed to provide the set of primary keys.



Settings (1)

Configure settings for the selected source data. You can add or delete source data in the source details section above.

Selected source data	Primary keys	Partition keys
/sap/opu/odata/sap/ZCUSTOMER_ATTR_1_SRV/EntityOf0CUSTOMER_ATTR	No primary keys selected	Default partition key

Configuring an Amazon DynamoDB source

To access data from your source Amazon DynamoDB table, AWS Glue requires access to describe the table and export data from it. Amazon DynamoDB recently introduced a feature which allows configuring a Resource Based Access (RBAC) policy.

The following example Resource Based Access (RBAC) Policy uses a wild card (*) for integration:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "1111",
      "Effect": "Allow",
      "Principal": {
        "Service": "glue.amazonaws.com"
      },
      "Resource": "*",
      "Action": [
        "dynamodb:ExportTableToPointInTime",
        "dynamodb:DescribeTable",
        "dynamodb:DescribeExport"
      ],
      "Condition": {
        "StringEquals": {
          "aws:SourceAccount": "111122223333"
        },
        "ArnLike": {
          "aws:SourceArn": "arn:aws:glue:us-east-1:111122223333:integration:*"
        }
      }
    }
  ]
}
```

1. For the DynamoDB that you want to replicate, paste the above RBAC policy template into **Resource-based policy for table** and fill in the fields.
2. If you want to make the policy restrictive, you must update the policy after creating the integration and specify the full `integrationArn` and use the `StringEquals` condition instead of `StringLike`.
3. Make sure that Point-in-time recovery (PITR) is enabled for the DynamoDB table.
4. Make sure that you add `Describe Export` to the Resource Based Access (RBAC) policy.

You can also add the RBAC policy to the table using the following command:

```
aws dynamodb put-resource-policy \
--resource-arn arn:aws:dynamodb:<region>:<account-id>:table/<ddb-table-name> \
--policy file://resource-policy-with-condition.json \
--region <region>
```

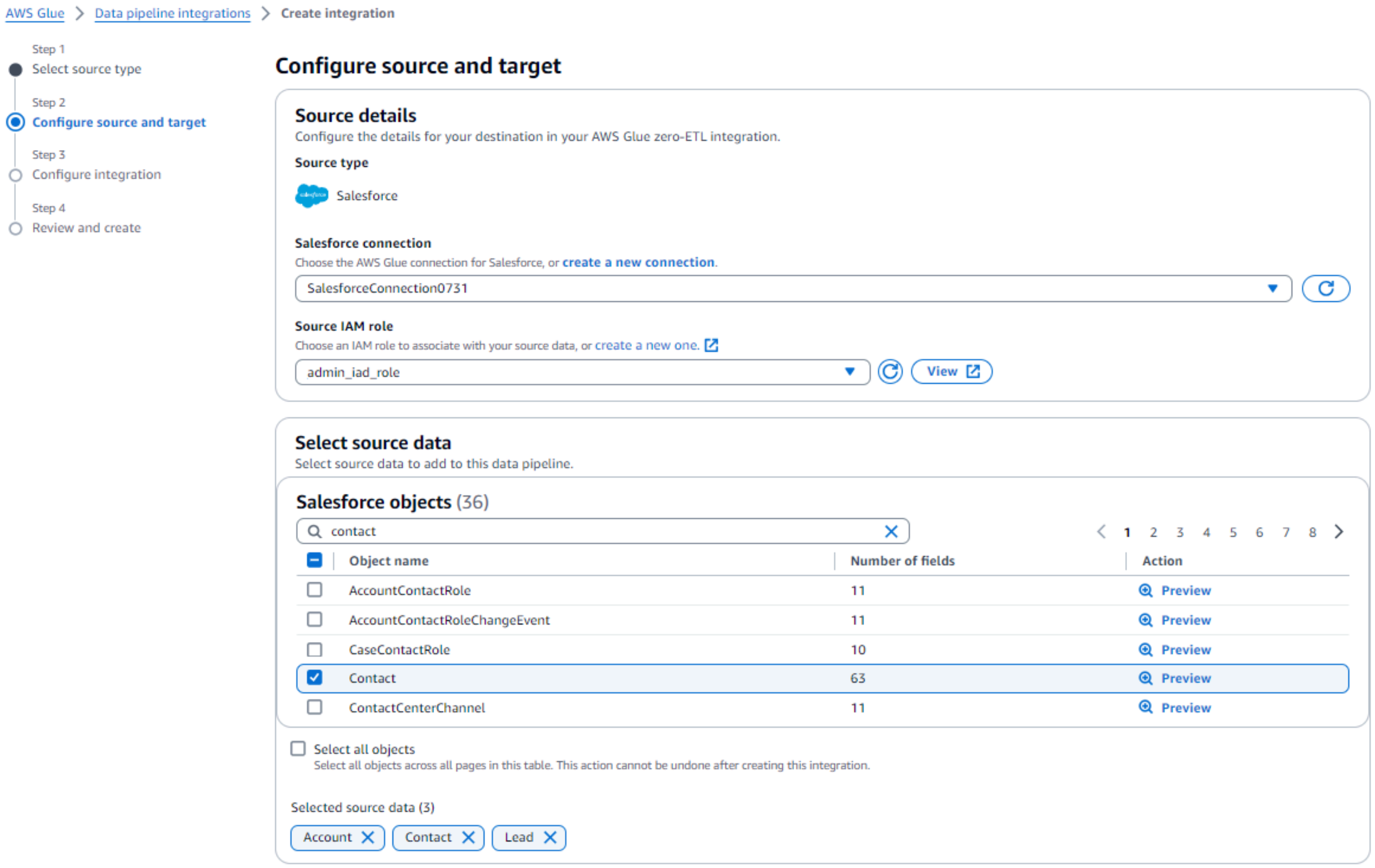
To verify that policy is applied correctly, use the following command to get the resource policy for a table:

```
aws dynamodb get-resource-policy \
--resource-arn arn:aws:dynamodb:<region>:<account-id>:table/<ddb-table-name>
--region <region>
```

Configuring a Salesforce source

To create a connection for an Salesforce source, see [Connecting to Salesforce](#).

Once you've created the connection, you can specify the source data to replicate.


AWS Glue > Data pipeline integrations > Create integration
Step 1
Select source type
Step 2
Configure source and target
Step 3
Configure integration
Step 4
Review and create

Configure source and target

Source details

Configure the details for your destination in your AWS Glue zero-ETL integration.

Source type: Salesforce

Salesforce connection: SalesforceConnection0731

Choose the AWS Glue connection for Salesforce, or create a new connection.

Source IAM role: admin_iad_role

Choose an IAM role to associate with your source data, or create a new one.

Select source data

Select source data to add to this data pipeline.

Salesforce objects (36)

contact

☐	Object name	Number of fields	Action
☐	AccountContactRole	11	Preview
☐	AccountContactRoleChangeEvent	11	Preview
☐	CaseContactRole	10	Preview
☑	Contact	63	Preview
☐	ContactCenterChannel	11	Preview

☐ Select all objects

Select all objects across all pages in this table. This action cannot be undone after creating this integration.

Selected source data (3)

Account ✕ Contact ✕ Lead ✕

Using your zero-ETL integration you can perform DDL operations for supported entities. For a list of entities which are not supported, see [Unsupported entities and fields for Salesforce](#).

Configuring a Salesforce Marketing Cloud Account Engagement source

To create a connection for an Salesforce Marketing Cloud Account Engagement source, see Connecting to Salesforce Marketing Cloud Account Engagement.

Using your zero-ETL integration you can perform DDL operations for the following supported entities:

Entity label	Entity name
Campaign	campaign
List	list
Dynamic Content	dynamic-content
List Membership	list-membership
Prospect	prospect
User	user
EmailTemplate	email-template
EngagementStudioProgram	engagement-studio-program
Landing Page	landing-page
List Email	list-email

Additional Salesforce Configuration

Salesforce Zero-ETL needs Lake Formation permission on Glue database, otherwise it will get `IngestionFailed` from the Log with the following error:

```
"errorMessage": "Insufficient lake formation permissions on Target Glue database."
```

Configuring an SAP OData source

To create a connection for an SAP OData source, see Connecting to SAP OData.

The SAP OData connector in a zero-ETL integration doesn't support entities starting with `EntityOf`.

Configuring a ServiceNow source

To create a connection for a ServiceNow source, see [Connecting to ServiceNow](#).

Configuring a Zendesk source

To create a connection for a Zendesk source, see [Connecting to Zendesk](#).

Using your zero-ETL integration you can perform the following DDL operations for supported entities:

Entity label	Entity name	Create supported	Update supported	Delete supported
Tickets	tickets	Y	Y	Y
User	users	Y	Y	Y
Satisfaction Rating	satisfaction-rating	Y	Y	N
Articles	articles	Y	Y	N
Organization	organizations	Y	Y	Y
Calls	calls	Y	Y	N
Call Legs	legs	Y	Y	N

Configuring a Zoho CRM source

To create a connection for a Zoho CRM source, see [Connecting to Zoho CRM](#).

Using your zero-ETL integration you can perform the following DDL operations for supported entities:

Entity label	Entity name	DML-Insert supported	DML-Modify supported	DML-Delete supported	DDL-Insert supported	DDL-Modify supported	DDL-Delete supported
Leads	lead	Y	Y	Y	Y	Y	Y
Accounts	account	Y	Y	Y	Y	Y	Y
Contacts	contact	Y	Y	Y	Y	Y	Y
Campaigns	campaign	Y	Y	Y	Y	Y	Y
Tasks	task	Y	Y	Y	Y	Y	Y
Events	event	Y	Y	Y	Y	Y	Y
Calls	call	Y	Y	Y	Y	Y	Y
Solutions	solution	Y	Y	Y	Y	Y	Y
Products	product	Y	Y	Y	Y	Y	Y
Vendors	vendor	Y	Y	Y	Y	Y	Y
Quotes	quote	Y	Y	Y	Y	Y	Y
Sales Orders	sales-order	Y	Y	Y	Y	Y	Y
Purchase Orders	purchase-order	Y	Y	Y	Y	Y	Y
Invoices	invoice	Y	Y	Y	Y	Y	Y
Cases	case	Y	Y	Y	Y	Y	Y
Price Books	price-book	Y	Y	Y	Y	Y	Y

Configuring a Facebook Ads source

To create a connection for a Facebook Ads source, see [Connecting to Facebook Ads](#).

Using your zero-ETL integration you can perform the following DDL operations for supported entities:

Entity label	Entity name	Create supported	Update supported	Delete supported
Adset	*/adsets	Y	Y	Y
Campaign	*/campaigns	Y	Y	Y
Ads	*/ads	Y	Y	Y

Configuring an Instagram Ads source

To create a connection for an Instagram Ads source, see [Connecting to Instagram Ads](#).

Using your zero-ETL integration you can perform the following DDL operations for supported entities:

Entity name	Create supported	Update supported	Delete supported
*/adsets	Y	Y	Y
*/campaigns	Y	Y	Y
*/ads	Y	Y	Y

Unsupported entities and fields for Salesforce

The following Salesforce entities or fields are unsupported for use in a zero-ETL integration with a Salesforce source.

```
AccountChangeEvent, AccountContactRoleChangeEvent, AccountHistory, AccountShare,
 ActiveFeatureLicenseMetric, ActivePermSetLicenseMetric, ActiveProfileMetric,
 ActivityFieldHistory, amzsec__asi_Telemetry_Data_Store__ChangeEvent,
amzsec__asi_Telemetry_Data_Store__History, amzsec__asi_Telemetry_Data_Store__Share,
amzsec__asi_Telemetry_Job_Log__ChangeEvent, amzsec__asi_Telemetry_Job_Log__History,
amzsec__asi_Telemetry_Job_Log__Share, amzsec__asi_Telemetry_Requirement__ChangeEvent,
amzsec__asi_Telemetry_Requirement__History, amzsec__asi_Telemetry_Requirement__Share,
ApexClass, ApexComponent, ApexLog, ApexPage, ApexTestQueueItem, ApexTestResult,
ApexTrigger, AssetChangeEvent, AssetHistory, AssetRelationshipHistory,
AssetShare, AssignmentRule, AssociatedLocationHistory, AsyncApexJob,
AuditTrailFileExportShare, AuthorizationFormConsentChangeEvent,
AuthorizationFormConsentHistory, AuthorizationFormConsentShare,
AuthorizationFormDataUseHistory, AuthorizationFormDataUseShare,
AuthorizationFormHistory, AuthorizationFormShare, AuthorizationFormTextHistory,
AuthProvider, AuthSession, BatchJobHistory, BatchJobPartFailedRecordHistory,
BatchJobPartHistory, BatchJobShare, BrandTemplate, BriefcaseAssignmentChangeEvent,
BriefcaseDefinitionChangeEvent, BusinessBrandShare, BusinessHours,
BusinessProcess, CalcMatrixColumnRangeHistory, CalcProcStepRelationshipHistory,
CalculationMatrixColumnHistory, CalculationMatrixHistory, CalculationMatrixRowHistory,
CalculationMatrixShare, CalculationMatrixVersionHistory, CalculationProcedureHistory,
CalculationProcedureShare, CalculationProcedureStepHistory,
CalculationProcedureVariableHistory, CalculationProcedureVersionHistory, Calendar,
CalendarViewShare, CallCenter, CallCoachConfigModifyEvent, CampaignChangeEvent,
CampaignHistory, CampaignMemberChangeEvent, CampaignMemberStatusChangeEvent,
CampaignShare, CaseChangeEvent, CaseHistory, CaseHistory2 CaseHistory2ChangeEvent,
CaseRelatedIssueChangeEvent, CaseRelatedIssueHistory, CaseShare, CaseStatus,
CaseTeamMember, CaseTeamRole, CaseTeamTemplate, CaseTeamTemplateMember,
CaseTeamTemplateRecord, CategoryNode, ChangeRequestChangeEvent, ChangeRequestHistory,
ChangeRequestRelatedIssueChangeEvent, ChangeRequestRelatedIssueHistory,
ChangeRequestRelatedItemChangeEvent, ChangeRequestRelatedItemHistory,
ChangeRequestShare, ChatRetirementRdyMetrics, ChatterActivity, ClientBrowser,
CollaborationGroup, CollaborationGroupMember, CollaborationGroupMemberRequest,
CollaborationInvitation, CommSubscriptionChannelTypeHistory,
CommSubscriptionChannelTypeShare, CommSubscriptionConsentChangeEvent,
CommSubscriptionConsentHistory, CommSubscriptionConsentShare, CommSubscriptionHistory,
CommSubscriptionShare, CommSubscriptionTimingHistory, Community, ConnectedApplication,
ContactChangeEvent, ContactHistory, ContactPointAddressChangeEvent,
ContactPointAddressHistory, ContactPointAddressShare, ContactPointConsentChangeEvent,
ContactPointConsentHistory, ContactPointConsentShare, ContactPointEmailChangeEvent,
ContactPointEmailHistory, ContactPointEmailShare, ContactPointPhoneChangeEvent,
ContactPointPhoneHistory, ContactPointPhoneShare, ContactPointTypeConsentChangeEvent,
ContactPointTypeConsentHistory, ContactPointTypeConsentShare, ContactRequestShare,
ContactShare, ContentDocumentChangeEvent, ContentDocumentHistory,
ContentDocumentLink, ContentDocumentLinkChangeEvent, ContentDocumentSubscription,
ContentFolderItem, ContentFolderLink, ContentFolderMember, ContentNote,
ContentNotification, ContentTagSubscription, ContentUserSubscription,
ContentVersionChangeEvent, ContentVersionComment, ContentVersionHistory,
ContentVersionRating, ContentWorkspace, ContentWorkspaceMember,
ContentWorkspacePermission, ContentWorkspaceSubscription, ContractChangeEvent,
ContractHistory, ContractLineItemChangeEvent, ContractLineItemHistory,
ContractStatus, Conversation, ConversationParticipant, CronJobDetail,
CronTrigger, CustomBrand, CustomBrandAsset, CustomerShare, CustomHttpHeader,
DashboardComponent, DataUseLegalBasisHistory, DataUseLegalBasisShare,
DataUsePurposeHistory, DataUsePurposeShare, DecisionTableRecordset, DeleteEvent,
DocumentAttachmentMap, Domain, DomainSite, DTRecordsetReplicaShare, EmailBounceEvent,
EmailMessageChangeEvent, EmailServicesAddress, EmailServicesFunction,
EmailTemplate, EmailTemplateChangeEvent, EngagementAttendeeChangeEvent,
EngagementAttendeeHistory, EngagementChannelTypeHistory, EngagementChannelTypeShare,
EngagementInteractionChangeEvent, EngagementInteractionHistory,
EngagementInteractionShare, EngagementInterface, EngagementTopicChangeEvent,
EngagementTopicHistory, EntitlementChangeEvent, EntitlementHistory,
EntitlementTemplate, EntityMilestoneHistory, EntitySubscription, EventChangeEvent,
EventRelationChangeEvent, EventRelayConfigChangeEvent, ExpressionSetHistory,
ExpressionSetShare, ExpressionSetVersionHistory, ExternalEventMappingShare,
FeedAttachment, FeedLike, FeedPollChoice, FeedPollVote, FeedSignal,
FieldPermissions, FieldSecurityClassification, FiscalYearSettings,
FlowInterviewLogShare, FlowInterviewShare, FlowOrchestrationEvent,
FlowOrchestrationInstanceShare, FlowOrchestrationStageInstanceShare,
FlowOrchestrationStepInstanceShare, FlowOrchestrationWorkItemShare,
FlowRecordShare, FlowRecordVersionChangeEvent, FlowTestResultShare, Folder,
Group, GroupMember, Holiday, IdeaComment, IdpEventLog, ImageHistory, ImageShare,
IncidentChangeEvent, IncidentHistory, IncidentRelatedItemChangeEvent,
IncidentRelatedItemHistory, IncidentShare, IndividualChangeEvent,
IndividualHistory, IndividualShare, KnowledgeableUser, LeadChangeEvent, LeadHistory,
LeadShare, LeadStatus, LightningExitByPageMetrics, LightningToggleMetrics,
LightningUsageByAppTypeMetrics, LightningUsageByBrowserMetrics,
LightningUsageByFlexiPageMetrics, LightningUsageByPageMetrics, ListEmailChangeEvent,
ListEmailShare, ListView, LocationChangeEvent, LocationHistory, LocationShare,
LocationTrustMeasureShare, LoginHistory, LoginIp, MacroChangeEvent,
MacroHistory, MacroInstructionChangeEvent, MacroShare, MacroUsageShare,
ManagedContentVariantChangeEvent, MessagingEndUserHistory, MessagingEndUserShare,
MessagingSessionHistory, MessagingSessionShare, MilestoneType, MLEngagementEvent,
ObjectPermissions, OpportunityChangeEvent, OpportunityContactRoleChangeEvent,
OpportunityFieldHistory, OpportunityLineItemChangeEvent, OpportunityShare,
OpportunityStage, OrderChangeEvent, OrderHistory, OrderItemChangeEvent,
OrderItemHistory, OrderShare, OrderStatus, Organization, OrgEmailAddressSecurity,
OrgWideEmailAddress, OutgoingEmail, OutgoingEmailRelation, PackageLicense,
PartnerRole, PartyConsentChangeEvent, PartyConsentHistory, PartyConsentShare,
Period, PermissionSet, PermissionSetAssignment, PermissionSetTabSetting,
Pricebook2ChangeEvent, Pricebook2History, PricebookEntryChangeEvent,
PricebookEntryHistory, PrivacyJobSessionShare, PrivacyObjectSessionShare,
PrivacyRTBFRequestHistory, PrivacyRTBFRequestShare, PrivacySessionRecordFailureShare,
ProblemChangeEvent, ProblemHistory, ProblemIncidentChangeEvent,
ProblemIncidentHistory, ProblemRelatedItemChangeEvent, ProblemRelatedItemHistory,
ProblemShare, ProcessDefinition, ProcessExceptionEvent, ProcessExceptionShare,
ProcessInstanceChangeEvent, ProcessInstanceStep, ProcessInstanceStepChangeEvent,
ProcessNode, Product2ChangeEvent, Product2History, ProductEntitlementTemplate,
Profile, ProfileSkillEndorsementHistory, ProfileSkillHistory,
ProfileSkillShare, ProfileSkillUserHistory, PromptActionShare, PromptErrorShare,
QueueSobject, QuickTextChangeEvent, QuickTextHistory, QuickTextShare,
QuickTextUsageShare, RecentlyViewed, RecommendationChangeEvent,
RecordActionHistory, RecordAlertHistory, RecordAlertShare, RecordType,
ScorecardShare, SellerHistory, SellerShare, ServiceContractChangeEvent,
ServiceContractHistory, ServiceContractShare, SetupAuditTrail, SetupEntityAccess,
SharingRecordCollectionShare, Site, SiteHistory, SiteRedirectMapping,
SocialPersonaHistory, SocialPostChangeEvent, SocialPostHistory, SocialPostShare,
SolutionHistory, SolutionStatus, StaticResource, StreamingChannelShare,
TableauHostMappingShare, TaskChangeEvent, TaskPriority, TaskStatus,
ThreatDetectionFeedback, TimelineObjectDefinitionChangeEvent, TodayGoalShare,
Topic, TopicUserEvent, Translation, User, UserAppMenuCustomizationShare,
UserChangeEvent, UserDefinedLabelAssignmentShare, UserDefinedLabelShare,
UserEmailPreferredPersonShare, UserLicense, UserLogin, UserPackageLicense,
UserPreference, UserPrioritizedRecordShare, UserProvisioningRequestShare,
UserRole, UserShare, VideoCallChangeEvent, VideoCallParticipantChangeEvent,
VideoCallRecordingChangeEvent, VideoCallShare, VisualforceAccessMetrics,
VoiceCallChangeEvent, VoiceCallRecordingChangeEvent, VoiceCallShare, Vote,
WebLink, WorkAccessShare, WorkBadgeDefinitionHistory, WorkBadgeDefinitionShare,
WorkOrderChangeEvent, WorkOrderHistory, WorkOrderLineItemChangeEvent,
WorkOrderLineItemHistory, WorkOrderLineItemStatus, WorkOrderShare, WorkOrderStatus,
WorkPlanChangeEvent, WorkPlanHistory, WorkPlanShare, WorkPlanTemplateChangeEvent,
WorkPlanTemplateEntryChangeEvent, WorkPlanTemplateEntryHistory,
WorkPlanTemplateHistory, WorkPlanTemplateShare, WorkStepChangeEvent, WorkStepHistory,
WorkStepStatus, WorkStepTemplateChangeEvent, WorkStepTemplateHistory,
WorkStepTemplateShare, WorkThanksShare
```

Unsupported entities and fields for ServiceNow

The following ServiceNow entities or fields are unsupported for use in a zero-ETL integration with a ServiceNow source.

```
ais_acl_overrides, ais_async_genius_result, ais_async_request, ais_connection,
 ais_genius_result_configuration_parameters, ais_partition_health,
 ais_partition_health_response, ais_publish_history, ais_relevancy_training_execution,
 ais_relevancy_training_staging, ais_search_profile_relevancy_model,
catalog_draft_entities, clone_log, clone_log0000, clone_log0001, clone_log0002,
clone_log0003, clone_log0004, clone_log0005, clone_log0006, clone_log0007,
cmdb_ie_context, cmdb_ie_log, cmdb_ie_run, cmdb_ire_partial_payloads_index,
cmdb_qb_result_base$par1, cmdb$par1, discovery_log, discovery_log0000,
discovery_log0001, discovery_log0002, discovery_log0003, discovery_log0004,
discovery_log0005, discovery_log0006, discovery_log0007, ecc_agent_log,
entitlement_data, entl_subscription_map, gs_entitlement_plugin_mapping,
ih_transaction_exclusion, import_log, import_log0000, import_log0001, import_log0002,
import_log0003, import_log0004, import_log0005, import_log0006, import_log0007,
jrobin_archive, jrobin_database, jrobin_datasource, jrobin_definition, jrobin_graph,
jrobin_graph_line, jrobin_graph_set, jrobin_graph_set_member, jrobin_shard,
jrobin_shard_location, license_role_discovery_run, logger_configuration_validation,
m2m_analytics_event_logger, m2m_user_consent_info, ml_artifact_object_store,
np$sys_gen_ai_filter_sample, np$sys_ui_element, np$sys_ui_list_element,
one_api_service_plan_feature_invocation, one_api_service_plan_invocation,
open_nlu_predict_log, open_nlu_predict_log0000, open_nlu_predict_log0001,
open_nlu_predict_log0002, open_nlu_predict_log0003, open_nlu_predict_log0004,
open_nlu_predict_log0005, open_nlu_predict_log0006, open_nlu_predict_log0007,
pa_diagnostic_log, pa_diagnostic_log0000, pa_diagnostic_log0001,
pa_diagnostic_log0002, pa_diagnostic_log0003, pa_diagnostic_log0004,
pa_diagnostic_log0005, pa_diagnostic_log0006, pa_diagnostic_log0007, pa_favorites,
pa_job_log_rows, pa_job_log_rows0000, pa_job_log_rows0001, pa_job_log_rows0002,
pa_job_log_rows0003, pa_job_log_rows0004, pa_job_log_rows0005, pa_job_log_rows0006,
pa_job_log_rows0007, pa_migration_ignored_scores, pa_scores_l1, pa_scores_l2,
pa_scores_migration_groups, par_dashboard_conversion_backup, promin_log,
promin_log0000, promin_log0001, promin_log0002, promin_log0003, promin_log0004,
promin_log0005, promin_log0006, promin_log0007, promin_request_object,
proposed_change_verification_log, proposed_change_verification_log0000,
proposed_change_verification_log0001, proposed_change_verification_log0002,
proposed_change_verification_log0003, proposed_change_verification_log0004,
proposed_change_verification_log0005, proposed_change_verification_log0006,
proposed_change_verification_log0007, protected_table_log, protected_table_log0000,
protected_table_log0001, protected_table_log0002, protected_table_log0003,
protected_table_log0004, protected_table_log0005, protected_table_log0006,
protected_table_log0007, pwd_history, qb_query_results, scan_log,
scan_log0000, scan_log0001, scan_log0002, scan_log0003, scan_log0004,
scan_log0005, scan_log0006, scan_log0007, schema_validator_error,
sla_repair_log_entry, sla_repair_log_entry0000, sla_repair_log_entry0001,
sla_repair_log_entry0002, sla_repair_log_entry0003, sla_repair_log_entry0004,
sla_repair_log_entry0005, sla_repair_log_entry0006, sla_repair_log_entry0007,
sla_repair_log_message, sla_repair_log_message0000, sla_repair_log_message0001,
sla_repair_log_message0002, sla_repair_log_message0003, sla_repair_log_message0004,
sla_repair_log_message0005, sla_repair_log_message0006, sla_repair_log_message0007,
sn_bm_client_activity, sn_ci_analytics_st_actionable_notifs,
sn_ci_analytics_st_conv_completion_by_cat, sn_ci_analytics_st_conv_dynamic_property,
sn_ci_analytics_st_conversation, sn_ci_analytics_st_count_by_date,
sn_ci_analytics_st_event_occurrence, sn_ci_analytics_st_event_property_value_trend,
sn_ci_analytics_st_issue_auto_resolution, sn_ci_analytics_st_no_clicks,
sn_ci_analytics_st_no_results, sn_ci_analytics_st_property_summary_by_event,
sn_ci_analytics_st_session_count_per_locale, sn_ci_analytics_st_session_duration,
sn_ci_analytics_st_spokes_usage, sn_ci_analytics_st_topic_execution_stats,
sn_ci_analytics_st_topic_occurrence, sn_ci_analytics_st_trending_content,
sn_ci_analytics_st_trending_queries, sn_ci_analytics_st_users,
sn_cs_plugin_signatures, sn_cs_telemetry_log, sn_dfc_application, sn_dfc_product,
sn_employee_position, sn_entitlement_st_subscription_application_family,
sn_entitlement_st_subscription_application_users, sn_hr_sp_st_relevant_for_you,
sn_instance_clone_log, sn_instance_clone_log0000, sn_instance_clone_log0001,
sn_instance_clone_log0002, sn_instance_clone_log0003, sn_instance_clone_log0004,
sn_instance_clone_log0005, sn_instance_clone_log0006, sn_instance_clone_log0007,
sn_km_mr_st_kb_knowledge, sn_me_st_topic, sn_rf_conditional_definition,
sn_rf_evaluation_type, sn_rf_evaluation_type_input, sn_rf_recommendation_action,
sn_rf_recommendation_experience, sn_rf_recommendation_history,
sn_rf_recommendation_rule, sn_rf_record_display_configuration, sn_rf_trend_definition,
sn_sub_man_st_account_level_entitlement, sn_sub_man_st_gen_ai_metadata,
sn_sub_man_st_instance_used_assist_count, sn_sub_man_st_now_assist_creator_instances,
sn_sub_man_st_now_assists_aggregate, sn_sub_man_st_subscribed_groups,
sn_sub_man_st_subscription_insights, sn_sub_man_st_subscription_license_detail_metric,
sn_sub_man_st_unallocated_group_recommendation, sn_sub_man_st_unconfirmed_user_group,
sn_wn_user_app_activity, sn_wn_user_content_activity, snc_monitorable_item,
snpar_sched_export_v_scheduled_export_visualization, spotlight,
spotlight_audit, spotlight_copy_log_row, spotlight_copy_log_row0000,
spotlight_copy_log_row0001, spotlight_copy_log_row0002, spotlight_copy_log_row0003,
spotlight_copy_log_row0004, spotlight_copy_log_row0005, spotlight_copy_log_row0006,
spotlight_copy_log_row0007, spotlight_job_log_row, spotlight_job_log_row0000,
spotlight_job_log_row0001, spotlight_job_log_row0002, spotlight_job_log_row0003,
spotlight_job_log_row0004, spotlight_job_log_row0005, spotlight_job_log_row0006,
spotlight_job_log_row0007, st_dfc_performance_metric, st_license_detail_metric,
st_on_call_hour, st_sc_wizard_question, st_sys_catalog_items_and_variable_sets,
st_sys_design_system_icon, subscription_instance_stats, svc_container_config,
svc_environment_config, svc_layer_config, svc_model_assoc_ci,
svc_model_checkpoint_attr, svc_model_obj_cluster, svc_model_obj_constraint,
svc_model_obj_deployable, svc_model_obj_element, svc_model_obj_impact,
svc_model_obj_impactrule, svc_model_obj_package, svc_model_obj_path,
svc_model_obj_relation, svc_model_obj_service, sys_administrative_script_transaction,
sys_amb_message, sys_amb_processor, sys_analytics_batch_state,
sys_analytics_config, sys_analytics_data_points_error, sys_analytics_event,
sys_analytics_logger, sys_analytics_logger_field, sys_app_payload_loader_rule,
sys_app_payload_unloader_rule, sys_app_scan_payload, sys_app_scan_variable,
sys_app_scan_variable_type, sys_archive_destroy_log, sys_archive_destroy_run,
sys_archive_log, sys_archive_run, sys_atf_transaction_log, sys_attachment_doc,
sys_attachment_doc_v2, sys_attachment_soft_deleted, sys_audit, sys_audit_relation,
sys_auth_policy_api_allowed, sys_aw_registered_scripting_modal,
sys_cache_flush, sys_data_egress_source, sys_dm_delete_count,
sys_export_set_log, sys_export_set_log0000, sys_export_set_log0001,
sys_export_set_log0002, sys_export_set_log0003, sys_export_set_log0004,
sys_export_set_log0005, sys_export_set_log0006, sys_export_set_log0007,
sys_flow_compiled_flow, sys_flow_compiled_flow_chunk, sys_flow_context_chunk,
sys_flow_context_chunk_archive, sys_flow_context_inputs_chunk,
sys_flow_execution_history, sys_flow_log, sys_flow_log0000, sys_flow_log0001,
sys_flow_log0002, sys_flow_log0003, sys_flow_plan_context_binding,
sys_flow_report_doc, sys_flow_report_doc_chunk, sys_flow_report_doc_chunk_archive,
sys_flow_runtime_state_chunk, sys_flow_runtime_value_chunk,
sys_flow_subflow_plan_chunk, sys_flow_trigger_plan_chunk, sys_flow_val_listener,
sys_flow_value, sys_flow_value_chunk, sys_gen_ai_config_example,
sys_gen_ai_feature_mapping, sys_gen_ai_strategy_mapping, sys_gen_ai_usage_log,
sys_generative_ai_capability_definition, sys_generative_ai_log,
sys_generative_ai_response_validator, sys_generative_ai_validator,
sys_geo_routing, sys_geo_routing_config, sys_hop_token, sys_hub_action_plan_chunk,
sys_hub_snapshot_chunk, sys_journal_field, sys_journal_field_edit, sys_json_chunk,
sys_kaa_policy, sys_kaa_subidentity_assertion, sys_kaa_user_policy_mapping,
sys_mapplication, sys_mass_encryption_job, sys_notification_execution_log,
sys_notification_execution_log0000, sys_notification_execution_log0001,
sys_notification_execution_log0002, sys_notification_execution_log0003,
sys_notification_execution_log0004, sys_notification_execution_log0005,
sys_notification_execution_log0006, sys_notification_execution_log0007,
sys_nowmq_message, sys_nowmq_provider_param_definition, sys_orchestrator_action,
sys_pd_asset_configuration, sys_pd_context_chunk, sys_pd_context_log,
sys_pd_snapshot_chunk, sys_pd_trigger_license, sys_processing_framework_job,
sys_query_index_hint, sys_query_rewrite, sys_query_string_log,
sys_replication_queue, sys_replication_queue0, sys_replication_queue1,
sys_replication_queue2, sys_replication_queue3, sys_replication_queue4,
sys_replication_queue5, sys_replication_queue6, sys_replication_queue7,
sys_request_performance, sys_rollback_blacklisted, sys_rollback_conflict,
sys_rollback_incremental, sys_rollback_log, sys_rollback_log0000,
sys_rollback_log0001, sys_rollback_log0002, sys_rollback_log0003,
sys_rollback_log0004, sys_rollback_log0005, sys_rollback_log0006,
sys_rollback_log0007, sys_rollback_run, sys_rollback_schema_change,
sys_rollback_schema_conflict, sys_rollback_sequence, sys_scheduler_assignment,
sys_scheduler_memory_pressure_job_log, sys_script_adapter_rule,
sys_script_batch_adapter_rule, sys_search_source_filter, sys_service_authentication,
sys_signing_job, sys_suggestion_reader, sys_sync_history_review, sys_trend,
sys_unreferenced_preview, sys_unreferenced_record_rule, sys_upgrade_manifest,
sys_upgrade_state, sys_ux_asset_cache_buster, sys_ux_lib_component_prop,
sys_ux_lib_presource, sys_ux_page_action, sys_ux_page_action_binding,
sysevent_queue_runtime, syslog, syslog_app_scope0000, syslog_app_scope0001,
syslog_app_scope0002, syslog_app_scope0003, syslog_app_scope0004,
syslog_app_scope0005, syslog_app_scope0006, syslog_app_scope0007, syslog_email0000,
syslog_email0001, syslog_email0002, syslog_email0003, syslog_email0004,
syslog_email0005, syslog_email0006, syslog_email0007, syslog_transaction,
syslog_transaction0000, syslog_transaction0001, syslog_transaction0002,
syslog_transaction0003, syslog_transaction0004, syslog_transaction0005,
syslog_transaction0006, syslog_transaction0007, syslog0000, syslog0001, syslog0002,
syslog0003, syslog0004, syslog0005, syslog0006, syslog0007, ts_attachment,
ts_c_1_0, ts_c_1_1, ts_c_1_2, ts_c_1_3, ts_c_1_4, ts_c_1_5, ts_c_1_6, ts_c_1_7,
ts_c_1_8, ts_c_1_9, ts_c_10_0, ts_c_10_1, ts_c_10_2, ts_c_10_3, ts_c_10_4,
ts_c_10_5, ts_c_10_6, ts_c_10_7, ts_c_10_8, ts_c_10_9, ts_c_11_0, ts_c_11_1,
ts_c_11_2, ts_c_11_3, ts_c_11_4, ts_c_11_5, ts_c_11_6, ts_c_11_7, ts_c_11_8,
ts_c_11_9, ts_c_2_0, ts_c_2_1, ts_c_2_2, ts_c_2_3, ts_c_2_4, ts_c_2_5, ts_c_2_6,
ts_c_2_7, ts_c_2_8, ts_c_2_9, ts_c_3_0, ts_c_3_1, ts_c_3_2, ts_c_3_3, ts_c_3_4,
ts_c_3_5, ts_c_3_6, ts_c_3_7, ts_c_3_8, ts_c_3_9, ts_c_39_0, ts_c_39_1, ts_c_39_2,
ts_c_39_3, ts_c_39_4, ts_c_39_5, ts_c_39_6, ts_c_39_7, ts_c_39_8, ts_c_39_9,
ts_c_40_0, ts_c_40_1, ts_c_40_2, ts_c_40_3, ts_c_40_4, ts_c_40_5, ts_c_40_6,
ts_c_40_7, ts_c_40_8, ts_c_40_9, ts_c_41_0, ts_c_41_1, ts_c_41_2, ts_c_41_3,
ts_c_41_4, ts_c_41_5, ts_c_41_6, ts_c_41_7, ts_c_41_8, ts_c_41_9, ts_c_42_0,
ts_c_42_1, ts_c_42_2, ts_c_42_3, ts_c_42_4, ts_c_42_5, ts_c_42_6, ts_c_42_7,
ts_c_42_8, ts_c_42_9, ts_c_43_0, ts_c_43_1, ts_c_43_2, ts_c_43_3, ts_c_43_4,
ts_c_43_5, ts_c_43_6, ts_c_43_7, ts_c_43_8, ts_c_43_9, ts_c_44_0, ts_c_44_1,
ts_c_44_2, ts_c_44_3, ts_c_44_4, ts_c_44_5, ts_c_44_6, ts_c_44_7, ts_c_44_8,
ts_c_44_9, ts_c_45_0, ts_c_45_1, ts_c_45_2, ts_c_45_3, ts_c_45_4, ts_c_45_5,
ts_c_45_6, ts_c_45_7, ts_c_45_8, ts_c_45_9, ts_c_6_0, ts_c_6_1, ts_c_6_2,
ts_c_6_3, ts_c_6_4, ts_c_6_5, ts_c_6_6, ts_c_6_7, ts_c_6_8, ts_c_6_9, ts_c_7_0,
ts_c_7_1, ts_c_7_2, ts_c_7_3, ts_c_7_4, ts_c_7_5, ts_c_7_6, ts_c_7_7, ts_c_7_8,
ts_c_7_9, ts_c_8_0, ts_c_8_1, ts_c_8_2, ts_c_8_3, ts_c_8_4, ts_c_8_5, ts_c_8_6,
ts_c_8_7, ts_c_8_8, ts_c_8_9, ts_c_attachment, ts_chain, ts_deleted_doc,
ts_document, ts_field, ts_index_stats, ts_phrase, ts_search_stats, ts_v4_attachment,
ts_word, ua_app_metadata, ua_audit_stats, ua_extra_page, ua_monitor_property,
ua_monitor_property_audit, ua_shared_service, ua_sn_table_inventory, ua_sp_known_bot,
ua_upload_log, v_ais_result_improvement_rule_condition_builder_values,
v_cluster_nodes, v_cxs_search_resource, v_db_index, v_db_trigger, v_file_load_order,
v_interaction_context, v_iostats, v_mysql_proclist, v_mysql_status, v_mysql_variables,
v_on_call_report_cache, v_pa_par_combined_dashboard, v_pd_activity_condition_to_run,
v_pd_activity_start_rule_with_condition, v_pd_lane_condition_to_run,
v_sql_debug, v_st_kb_category, v_st_kb_most_viewed, v_st_kb_recently_viewed,
v_st_km_genai_mra_similar_task, v_st_popular_item, v_st_recent_item, v_st_sc_cat_item,
v_st_sc_catalog, v_st_sc_category, validator_run_summary, vtb_card_history,
wf_log, wf_log0000, wf_log0001, wf_log0002, wf_log0003, wf_log0004, wf_log0005,
wf_log0006, wf_log0007, spotlight_criteria, pa_snapshots, pa_widget_indicators,
pa_widgets, sn_cim_register, global, gsw_change_log, gsw_content, gsw_content_group,
gsw_content_information, gsw_status_of_content, multi_factor_browser_fingerprint,
multi_factor_criteria, pa_filters, password_policy, plan_execution,
plan_mysql, plan_oracle, plan_postgres, sc_rest_api_without_access_policy,
sn_actsub_activity, sn_actsub_activity_fanout, sn_actsub_activity_stream,
sn_actsub_activity_type, sn_actsub_atype_attributes, sn_actsub_atype_notif_pref,
sn_actsub_module, sn_actsub_notif_object, sn_actsub_subobject_stream,
sn_actsub_subscribable_object, sn_actsub_subscription_notif_pref,
sn_actsub_user_stream, sn_appclient_store_outbound_http_quota,
sn_appcreator_app_template, sn_critical_update, sn_docker_spoke_images,
sn_employee_app, sn_employee_app_access, sn_employee_app_access_criteria,
sn_entitlement_genai_assist_counts, sn_entitlement_genai_creator_user_counts,
sn_entitlement_genai_creator_users, sn_mif_instance, sn_mif_sync_data,
sn_mif_sync_status, sn_mif_table_registration, sn_mif_trust_config,
sn_vsc_best_practice_configurations, sn_vsc_best_practice_goals,
sn_vsc_changed_hardening_settings, sn_vsc_changed_scan_findings,
sn_vsc_check_security_area, sn_vsc_elevation_event, sn_vsc_event,
sn_vsc_export_event, sn_vsc_export_setting, sn_vsc_harc_compliance_status_lookup,
sn_vsc_hardening_compliance_scores, sn_vsc_impersonation_event,
sn_vsc_instance_hardening_settings, sn_vsc_login_event,
sn_vsc_scan_comparisons, sn_vsc_scan_summary, sn_vsc_security_check_categories,
sn_vsc_security_check_configurations, sn_vsc_security_configuration_groups,
sn_vsc_security_privacy_capabilities, sn_vsc_updated_settings,
sn_vsc_user_comparisons, sys_app_hash_inventory, sys_coalesce_strategy_deferred,
sys_flow_secure_data, sys_formula_function, sys_geocoding_request,
sys_global_file_hash, sys_import_set_row, sys_index, sys_index_explain,
sys_installation_schedule, sys_installation_schedule_item, sys_offline_app,
sys_package, sys_package_dependency_item, sys_package_dependency_m2m,
sys_plugins, sys_querystat, sys_reap_package, sys_scoped_plugin,
sys_stage_storage_alias, sys_storage_alias, sys_storage_table_alias, sys_store_app,
sys_table_partition, sys_upgrade_history_log, sys_user_public_credential,
sys_webauthn_authentication_request, sys_webauthn_registration_request,
syslog_app_scope, syslog_email, syslog_page_timing, ua_instance_state_config,
v_expression_cache, v_private_cache, v_shared_cache, sys_amb_message0002,
sys_amb_message0004, sys_amb_message0005, sys_metadata, v_par_unified_report_viz
```

Configuring a zero-ETL integration target

There are several options offered by AWS when configuring a target for a zero-ETL integration. The target may be an encrypted Amazon Redshift data warehouse or an Amazon SageMaker Lakehouse catalog.

Before selecting the target for the zero-ETL integration, you need to configure one of the following target resources.

The configuration options for a target in a zero-ETL integration include:

- An Amazon SageMaker Lakehouse catalog and database configured with regular Amazon S3 storage. See Configuring an Amazon SageMaker Lakehouse catalog with regular S3 storage.
- An Amazon SageMaker Lakehouse catalog configured with Amazon S3 Tables bucket. See Configuring Amazon S3 tables as a target.
- An Amazon SageMaker Lakehouse catalog configured with Amazon Redshift managed storage. See Configuring an Amazon SageMaker Lakehouse catalog with Amazon Redshift managed storage.
- An Amazon Redshift data warehouse identified by a Redshift namespace. See Configuring an Amazon Redshift data warehouse target.

> **Note**
>
> You cannot modify the target of a zero-ETL integration after creation.

Configuring an Amazon SageMaker Lakehouse catalog with regular S3 storage

This section describes the prerequisites and setup steps for configuring a regular Amazon S3 bucket as storage for your Amazon SageMaker Lakehouse catalog target in a zero-ETL integration.

Prerequisites for setting up an integration

Before creating a zero-ETL integration with an Amazon SageMaker Lakehouse catalog using regular S3 storage, you need to complete the following setup tasks:

1. Set up an AWS Glue database
2. Provide Catalog RBAC policy
3. Create target IAM role

After configuring the Amazon SageMaker Lakehouse catalog with regular Amazon S3 storage, you can proceed to Configuring the integration with your target to complete the integration setup.

Configuring Amazon S3 tables as a target

This section describes the prerequisites and setup steps for configuring Amazon S3 Tables as a target for your zero-ETL integration.

Prerequisites for setting up an integration

Before creating a zero-ETL integration with Amazon S3 Tables as a target, you need to complete the following setup tasks:

1. Setup Amazon S3 tables bucket
2. Provide Catalog RBAC policy
3. Create target IAM role

Setup Amazon S3 tables bucket

1. Create an S3 table bucket in your account by following the instructions at Getting started with Amazon S3 Tables.
2. Enable Analytics integrations with your S3-Table bucket by following these instructions: Integrating AWS services with Amazon S3 Tables.

Provide Catalog RBAC Policy

The following permissions must be added to the Catalog RBAC Policy to allow for integrations between source and Amazon S3 tables catalog target.

Target AWS Glue Catalog resource policy needs to include Glue Service permissions to AuthorizeInboundIntegration. Additionally, CreateInboundIntegration permission is required either on the source principal creating the Integration or in the target AWS Glue resource policy.

> **Note**
>
> For cross-account scenario, both source principal as well as target AWS Glue Catalog resource policy need to include glue:CreateInboundIntegration permissions on the resource.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Principal": {
        "AWS": [
         "arn:aws:iam::123456789012:user/Alice"
        ]
      },
      "Effect": "Allow",
      "Action": [
        "glue:CreateInboundIntegration"
      ],
      "Resource": [
       "arn:aws:glue:us-east-1:111122223333:catalog/<s3tablescatalog>/*"
      ],
      "Condition": {
        "StringLike": {
        "aws:SourceArn": "arn:aws:dynamodb:us-east-1:444455556666:table/<table-
name>"
        }
      }
    },
    {
      "Principal": {
        "Service": [
          "glue.amazonaws.com"
        ]
      },
      "Effect": "Allow",
      "Action": [
        "glue:AuthorizeInboundIntegration"
      ],
      "Resource": [
      "arn:aws:glue:us-east-1:111122223333:catalog/<s3tablescatalog>/*"
      ],
      "Condition": {
        "StringEquals": {
        "aws:SourceArn": "arn:aws:dynamodb:us-east-1:444455556666:table/<table-
name>"
        }
        }
      }
    ]
}
```

Note

Replace `<s3tablescatalog>` with the catalog name of your S3 tables.

Create target IAM Role

Create a target IAM role with the following permissions and trust relationships:

Example IAM policy:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Action": [
        "s3tables:ListTableBuckets",
        "s3tables:GetTableBucket",
        "s3tables:GetTableBucketEncryption",
        "s3tables:GetNamespace",
        "s3tables:CreateNamespace",
        "s3tables:ListNamespaces",
        "s3tables:CreateTable",
        "s3tables:GetTable",
        "s3tables:GetTableEncryption",
        "s3tables:ListTables",
        "s3tables:GetTableMetadataLocation",
        "s3tables:UpdateTableMetadataLocation",
        "s3tables:GetTableData",
        "s3tables:PutTableData"
      ],
      "Resource": "arn:aws:s3tables:us-east-1:111122223333:bucket/*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "cloudwatch:PutMetricData"
      ],
      "Resource": "*",
      "Condition": {
        "StringEquals": {
          "cloudwatch:namespace": "AWS/Glue/ZeroETL"
        }
      },
      "Effect": "Allow"
    },
    {
      "Action": [
        "logs:CreateLogGroup",
        "logs:CreateLogStream",
        "logs:PutLogEvents"
      ],
      "Resource": "*",
      "Effect": "Allow"
    }
  ]
}
```

Add the following trust policy in the Target IAM role to allow AWS Glue Service to assume it:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {
        "Service": "glue.amazonaws.com"
      },
      "Action": "sts:AssumeRole"
    }
  ]
}
```

Note

Make sure there is no explicit DENY statement for this target IAM role in the S3-Tables bucket resource policy. An explicit DENY would override any ALLOW permissions and prevent the integration from working properly.

Configuring an Amazon SageMaker Lakehouse catalog with Amazon Redshift managed storage

This section describes the prerequisites and setup steps for configuring an Amazon SageMaker Lakehouse catalog with Amazon Redshift managed storage (RMS) as a target for your zero-ETL integration.

Prerequisites for setting up an integration

Before creating a zero-ETL integration with an Amazon SageMaker Lakehouse catalog using Redshift managed storage, you need to complete the following setup tasks:

1. Set up an Amazon Redshift cluster or Serverless workgroup
2. Register the Amazon Redshift integration with Lake Formation
3. Create a managed catalog in Lake Formation
4. Configure IAM permissions

Setting up Amazon Redshift managed storage

To set up Amazon Redshiftmanaged storage for your zero-ETL integration:

1. Create or use an existing Amazon Redshift cluster or Serverless workgroup. Make sure the target Amazon Redshift workgroup or cluster has the `enable_case_sensitive_identifier` parameter turned on for the integration to be successful. For more information on enabling case sensitivity, see Turn on case sensitivity for your data warehouse in the Amazon Redshift management guide.
2. Register an integration from Redshift into the catalog in AWS Lake Formation. See Registering Amazon Redshift clusters and namespaces to the AWS Glue Data Catalog.
3. Create a federated or managed catalog in AWS Lake Formation. For more information, see:

  - Bringing Amazon Redshift data into the AWS Glue Data Catalog
  - Creating an Amazon Redshift managed catalog in the AWS Glue Data Catalog
4. Configure IAM permissions for the target role. The role needs permissions to access both Redshift and Lake Formation resources. At minimum, the role should have:
  - Permissions to access the Redshift cluster or workgroup
  - Permissions to access the Lake Formation catalog
  - Permissions to create and manage tables in the catalog
  - CloudWatch and CloudWatch Logs permissions for monitoring

After configuring the Amazon SageMaker Lakehouse catalog with Amazon Redshift managed storage, you can proceed to Configuring the integration with your target to complete the integration setup.

Configuring an Amazon Redshift data warehouse target

This section describes the prerequisites and setup steps for configuring an Amazon Redshift data warehouse as a target for your zero-ETL integration.

Prerequisites for setting up an integration

Before creating a zero-ETL integration with an Amazon Redshift data warehouse target, you need to complete the following setup tasks:

1. Set up an Amazon Redshift cluster or Serverless workgroup
2. Configure case sensitivity
3. Configure IAM permissions

Setting up the Amazon Redshift data warehouse

To set up an Amazon Redshift data warehouse for your zero-ETL integration:

1. Navigate to the Amazon Redshift console and click **Create cluster** or use an existing cluster. For Amazon Redshift Serverless, click **Create workgroup**.
2. If creating a new cluster, choose an appropriate cluster size and ensure your cluster is encrypted. For Serverless, configure the workgroup settings according to your requirements.

3. Make sure the target Amazon Redshift workgroup or cluster has the `enable_case_sensitive_identifier` parameter turned on for the integration to be successful. For more information on enabling case sensitivity, see Turn on case sensitivity for your data warehouse in the Amazon Redshift management guide.
4. Configure IAM permissions to allow the zero-ETL integration to access your Amazon Redshift data warehouse. You'll need to create an IAM role with the following permissions:
    - Permissions to access the Amazon Redshift cluster or workgroup
    - Permissions to create and manage databases and tables in Amazon Redshift
    - CloudWatch and Amazon CloudWatch Logs permissions for monitoring
5. After the Amazon Redshift workgroup or cluster setup is complete, you need to configure your data warehouse for zero-ETL integrations. See Getting started with zero-ETL integrations in the Amazon Redshift Management Guide for more information.

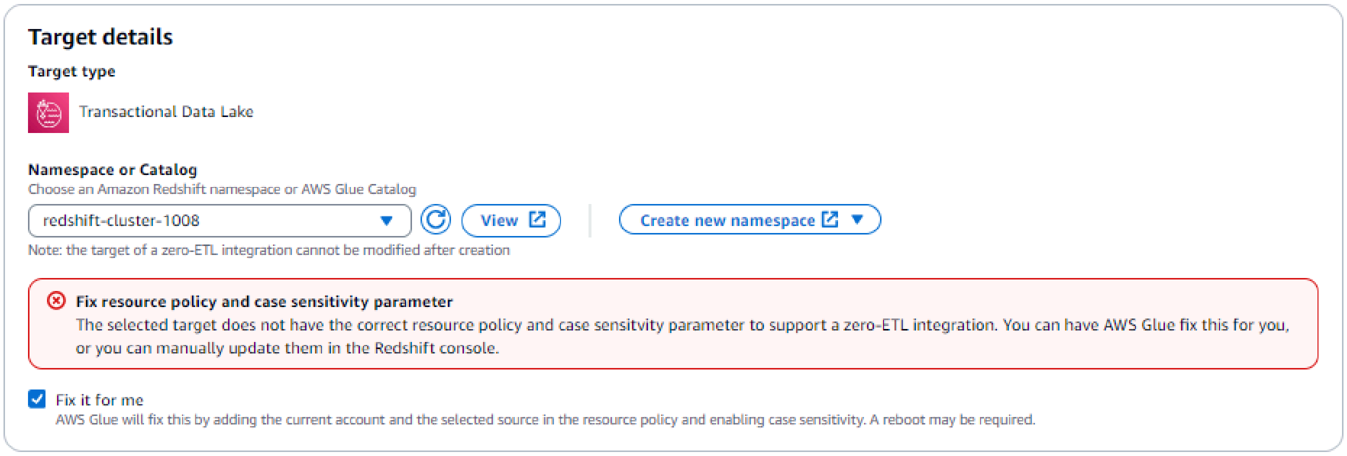
> **Note**
>
> When using a Amazon Redshift data warehouse as a target, the integration creates a schema in the specified database to store the replicated data. The schema name is derived from the integration name.

After configuring the Amazon Redshift data warehouse, you can proceed to Configuring the integration with your target to complete the integration setup.

Configuring the integration with your target

After you have configured your target resources and selected your connection and specified a source IAM role, follow these steps to complete the integration setup:

1. Specify the target you've configured in the previous steps.
2. Select the AWS Glue **Fix it for me** option. For the Amazon Redshift target, this will:
    - Apply an authorized service principal on the Amazon Redshift cluster or Serverless workgroup.
    - Apply an authorized AWS Glue source ARN to the Amazon Redshift cluster or Serverless workgroup.
    - Associate a new parameter group with `enable_case_sensitive_identifier = true`.



Target details

Target type

Transactional Data Lake

Namespace or Catalog

Choose an Amazon Redshift namespace or AWS Glue Catalog

redshift-cluster-1008

Note: the target of a zero-ETL integration cannot be modified after creation

Fix resource policy and case sensitivity parameter

The selected target does not have the correct resource policy and case sensitvity parameter to support a zero-ETL integration. You can have AWS Glue fix this for you, or you can manually update them in the Redshift console.

☑ Fix it for me

AWS Glue will fix this by adding the current account and the selected source in the resource policy and enabling case sensitivity. A reboot may be required.

3. Provide the integration name and choose **Create and launch Integration**.
4. Once your integration is in the active state, navigate to the integration details page and choose **Create a database from integration**.
5. Finally, you can navigate to the Redshift query editor, and connect to your database to validate the snapshot and incremental data.

> **Note**
>
> You can only use lowercase alphanumeric characters and underscores in the namespace or catalog name. This is different from what the AWS Glue Data Catalog allows to create a database with any name (including special characters).

Partition specification and schema unnesting guide

When working with NoSQL data sources like DynamoDB and SaaS applications, data often presents unique challenges for analytics:

1. Records within the same table may have different schema
2. Nested records within the same table can be represented differently
3. Complex nested structures like maps and arrays require transformation for efficient querying
4. Optimal data organization is needed to ensure query performance at scale

AWS Glue Zero-ETL integrations address these challenges through two powerful capabilities:

- **Schema Unnesting:** Automatically flattens complex nested data structures into analytics-friendly formats, with configurable levels of unnesting to balance between preserving data structure and optimizing for query simplicity.
- **Data Partitioning:** Organizes data into logical partitions based on specified columns or time-based dimensions, improving query performance and reducing costs by enabling partition pruning during query execution.

In order to query such data sources effectively, AWS Glue Zero-ETL provides out-of-the-box schema handling and partitioning schemes for source data being replicated in the target AWS Glue Database. You can configure schema unnesting and partitioning settings for each table through the CreateIntegrationTableProperty API, allowing for fine-tuned control over how data is structured and organized for analytics workloads.

Default unnesting & partitioning behavior

1. AWS Glue Zero-ETL defaults to FULL Unnest when no Unnesting options are provided for target table
2. AWS Glue Zero-ETL defaults to Bucket partitioning when no PartitionSpec are provided for target table

Schema unnesting

When integrating with analytics services through Zero-ETL, you can choose how nested structures are represented in the target tables. AWS Glue Zero-ETL provides schema unnesting options to flatten complex data structures into more analytics-friendly formats.

Unnesting options

When creating a Zero-ETL integration with a source, you can choose from the following unnesting options. These options correspond to specific enumeration values that you'll use when calling the CreateIntegrationTableProperty API.

No unnesting (default)

API value: `NO_UNNEST`

Preserves the original nested structure of DynamoDB items. Maps and lists are stored as structured columns in the target.

Best for: Preserving the exact structure of your DynamoDB data when your analytics tools can work with nested data.

Unnest one level

API value: TOP_LEVEL

Flattens the top level of nested maps into individual columns. List structures remain nested.

Best for: Balancing between data structure preservation and query simplicity when your DynamoDB items have a consistent schema.

Unnest all levels

API value: FULL

Recursively flattens all nested structures (maps and lists) into individual columns with dot notation for naming.

Best for: Maximizing query simplicity when working with deeply nested structures and analytics tools that prefer flat schemas.

> **Note**
>
> Full unnesting can lead to very wide tables with many columns if your DynamoDB data has variable or deeply nested structures.

Example Using unnesting options in the API

When configuring schema unnesting through the CreateIntegrationTableProperty API, specify the unnesting option in the UnnestSpec parameter:

```
aws glue create-integration-table-property
  --resource-arn "arn:aws:glue:us-east-1:123456789012:integration/my-integration"
  --table-name "my-table"
  --cli-input-json '{
      "TargetTableConfig": {
          "UnnestSpec": "FULL",
          "TargetTableName": "my-target-table",
      }
  }'
```

Unnesting examples

Consider a DynamoDB item with the following structure:

```
{
  "ProductId": "P12345",
  "ProductDetails": {
    "Name": "Smartphone",
    "Brand": "TechCo",
    "Specifications": {
      "Color": "Black",
      "Storage": "128GB"
    }
  },
  "Reviews": [
    {
      "Rating": 5,
      "Comment": "Great product!"
    },
    {
      "Rating": 4,
      "Comment": "Good value."
    }
  ]
}
```

No unnesting example

With no unnesting, the target table would have columns:

- ProductId (string)
- ProductDetails (struct)
- Reviews (array of structs)

Queries would need to use struct and array access patterns:

```
SELECT
  ProductId,
  ProductDetails.Name,
  ProductDetails.Specifications.Color
FROM product_table;
```

Unnest one level example

With one level unnesting, the target table would have columns:

- ProductId (string)
- ProductDetails_Name (string)
- ProductDetails_Brand (string)
- ProductDetails_Specifications (struct)
- Reviews (array of structs)

Queries would be simplified for the first level:

```
SELECT
  ProductId,
  ProductDetails_Name,
  ProductDetails_Specifications.Color
FROM product_table;
```

Unnest all levels example

With all levels unnesting, the target table would have columns and values:

Column Name (Type)	Value
ProductId (string)	P12345
ProductDetails_Name (string)	Smartphone
ProductDetails_Brand (string)	TechCo
ProductDetails_Specifications_Color (string)	Black
ProductDetails_Specifications_Storage (string)	128GB
Reviews_0_Rating (number)	5
Reviews_0_Comment (string)	Great product!
Reviews_1_Rating (number)	4

Column Name (Type)	Value
Reviews_1_Comment (string)	Good value.

Queries would be fully flattened:

```
SELECT
  ProductId,
  ProductDetails_Name,
  ProductDetails_Specifications_Color
FROM product_table;
```

Data partitioning

What is data partitioning?

Data partitioning is a technique that divides large datasets into smaller, more manageable segments called partitions. In the context of AWS Glue Zero-ETL integrations, partitioning organizes your data in the target location based on specific column values or transformations of those values.

Benefits of data partitioning

Effective data partitioning provides several key benefits for analytics workloads:

- **Improved query performance:** Queries can skip irrelevant partitions (partition pruning), reducing the amount of data that needs to be scanned.
- **Reduced costs:** By scanning less data, you can lower compute and I/O costs for your analytics queries.
- **Better scalability:** Partitioning allows parallel processing of data segments, enabling more efficient scaling of analytics workloads.
- **Simplified data lifecycle management:** You can manage retention policies at the partition level, making it easier to archive or delete older data.

Key partitioning concepts

Partition columns

Columns in your data that are used to determine how records are organized into partitions. Effective partition columns should align with common query patterns and have appropriate cardinality.

Partition functions

Transformations applied to partition column values to create the actual partition boundaries. Examples include identity (using the raw value) and time-based functions (year, month, day, hour).

Partition pruning

The process where the query engine identifies and skips partitions that don't contain relevant data for a query, significantly improving performance.

Partition granularity

The level of detail at which data is partitioned. Finer granularity (more partitions) can improve query performance but may increase metadata overhead. Coarser granularity (fewer partitions) reduces metadata overhead but may result in scanning more data than necessary.

Partitioning in AWS Glue Zero-ETL integrations

AWS Glue Zero-ETL integrations use Apache Iceberg table format, which provides advanced partitioning capabilities. When you create a Zero-ETL integration, you can:

- Use default partitioning strategies optimized for your data source
- Define custom partitioning specifications tailored to your query patterns
- Apply transformations to partition columns (especially useful for timestamp-based partitioning)
- Combine multiple partition strategies for multi-level partitioning

Partitioning configurations are specified through the `CreateIntegrationTableProperty` API when setting up your Zero-ETL integration. Once configured, AWS Glue automatically applies these partitioning strategies to organize your data in the target location.

Partition specification API reference

Use the following parameters in the CreateIntegrationTableProperties API to configure partitioning:

PartitionSpec

An array of partition specifications that defines how data is partitioned in the target location.

```
{
  "partitionSpec": [
    {
      "fieldName": "timestamp_col",
      "functionSpec": "month",
      "conversionSpec": "epoch_milli"
    },
    {
      "fieldName": "category",
      "functionSpec": "identity"
    }
  ]
}
```

FieldName

A UTF-8 string (1-128 bytes) specifying the column name to use for partitioning.

FunctionSpec

Specifies the partitioning function. Valid values:

- `identity` - Uses source values directly without transformation
- `year` - Extracts the year from timestamp values (e.g., 2023)
- `month` - Extracts the month from timestamp values (e.g., 2023-01)
- `day` - Extracts the day from timestamp values (e.g., 2023-01-15)
- `hour` - Extracts the hour from timestamp values (e.g., 2023-01-15-14)

Note

Time-based functions (`year`, `month`, `day`, `hour`) require the `ConversionSpec` parameter to specify the source timestamp format.

ConversionSpec

A UTF-8 string that specifies the timestamp format of the source data. Valid values are:

- `epoch_sec` - Unix epoch timestamp in seconds
- `epoch_milli` - Unix epoch timestamp in milliseconds
- `iso` - ISO 8601 formatted timestamp

Partitioning strategies

Default partitioning

When no partition columns are specified, AWS Glue Zero-ETL applies default partitioning strategies optimized for your data source:

- **Primary key-based partitioning:** For sources with primary keys (like DynamoDB tables), AWS Glue Zero-ETL automatically partitions data using the primary key with bucketing to prevent partition explosion.

Default partitioning is designed to work well for common query patterns without requiring manual configuration. However, for specific query patterns or performance requirements, you may want to define custom partitioning strategies.

User-defined partitioning strategies

AWS Glue Zero-ETL allows you to define custom partitioning strategies using the `PartitionSpec` parameter. You can specify one or more partition columns and apply different partitioning functions to each column.

Identity partitioning uses the raw values from a column to create partitions. This strategy is useful for columns with low to medium cardinality, such as category, region, or status fields.

Example Identity partitioning example

```
{
  "partitionSpec": [
    {
      "fieldName": "category",
      "functionSpec": "identity"
    }
  ]
}
```

This creates separate partitions for each unique value in the "category" column.

> **Warning**
>
> Avoid using identity partitioning with high-cardinality columns (like primary keys or timestamps) as it can lead to partition explosion, which degrades performance and increases metadata overhead.

Time-based partitioning organizes data based on timestamp values at different granularities (year, month, day, or hour). This strategy is ideal for time-series data and enables efficient time-range queries.

When using time-based partitioning, AWS Glue Zero-ETL can automatically convert various timestamp formats to a standardized format before applying the partition function. This conversion is specified using the `ConversionSpec` parameter.

Example Time-based partitioning example

```
{
  "partitionSpec": [
    {
      "fieldName": "created_at",
      "functionSpec": "month",
      "conversionSpec": "epoch_milli"
    }
  ]
}
```

This partitions data by month based on the "created_at" column, which contains Unix epoch timestamps in milliseconds.

AWS Glue Zero-ETL supports the following time-based partition functions:

- **year:** Partitions data by year (e.g., 2023, 2024)
- **month:** Partitions data by month (e.g., 2023-01, 2023-02)
- **day:** Partitions data by day (e.g., 2023-01-01, 2023-01-02)

- **hour:** Partitions data by hour (e.g., 2023-01-01-01, 2023-01-01-02)

AWS Glue Zero-ETL supports the following timestamp formats through the `ConversionSpec` parameter:

- **epoch_sec:** Unix epoch timestamps in seconds
- **epoch_milli:** Unix epoch timestamps in milliseconds
- **iso:** ISO 8601 formatted timestamps

> **Note**
>
> The original column values remain unchanged in your source data. AWS Glue only transforms partition column values to Timestamp Type in the target database table. The transformations only apply to the partitioning process.

Multi-level partitioning combines multiple partition strategies to create a hierarchical partitioning scheme. This is useful for optimizing different types of queries against the same dataset.

Example Multi-level partitioning example

```
{
  "partitionSpec": [
    {
      "fieldName": "created_at",
      "functionSpec": "month",
      "conversionSpec": "iso"
    },
    {
      "fieldName": "region",
      "functionSpec": "identity"
    }
  ]
}
```

This creates a two-level partitioning scheme: first by month (from the "created_at" column), then by region. This enables efficient queries that filter by date ranges, specific regions, or a combination of these dimensions.

When designing multi-level partitioning schemes, consider:

- Placing higher-selectivity columns first in the partition hierarchy
- Balancing partition granularity with the number of partitions
- Aligning the partitioning scheme with your most common query patterns

Best practices

Partition column selection

- Do not use high-cardinality columns with the `identity` partition function. Using high-cardinality columns with identity partitioning creates many small partitions, which can significantly degrade ingestion performance. High-cardinality columns may include:
  - Primary keys
  - Timestamp fields (such as `LastModifiedTimestamp`, `CreatedDate`)
  - System-generated timestamps
- Do not select multiple timestamp partitions on same column. For example:

```
"partitionSpec": [
      {"fieldName": "col1", "functionSpec": "year", "conversionSpec" :
 "epoch_milli"},
      {"fieldName": "col1", "functionSpec": "month", "conversionSpec" :
 "epoch_milli"},
      {"fieldName": "col1", "functionSpec": "day", "conversionSpec" : "epoch_milli"},
      {"fieldName": "col1", "functionSpec": "hour", "conversionSpec" : "epoch_milli"}
]
```

Partition FunctionSpec/ConversionSpec selection

- Specify the correct ConversionSpec (epoch_sec | epoch_milli | iso) that represents format of column values chosen for timestamp based partitioning when using timestamp-based partition functions. AWS Glue Zero-ETL uses this parameter to correctly transform source data into timestamp format before partitioning.
- Use appropriate granularity (year/month/day/hour) based on data volume.
- Consider timezone implications when using ISO timestamps. AWS Glue Zero-ETL populates all the record values of chosen timestamp column with UTC timezone.

Error handling

NEEDS_ATTENTION State

An integration enters the NEEDS_ATTENTION state when:

- Specified partition columns do not exist in the source
- Timestamp conversion fails for partition columns

Limitations

Partitioning limitations

- Partition specifications cannot be changed after an integration is created. To use a different partitioning strategy, you must create a new integration.
- The maximum number of partition columns is limited to 10.

Cross-account integration limitations

- When creating cross-account integrations, AWS Glue Console has a limitation where it doesn't invoke CreateIntegrationTableProperty API for configuring UnnestSpec and PartitionSpec for target AWS Glue tables that are hosted in the account where integration does not exist.

  Workaround: The CreateIntegrationTableProperty API must be invoked by CX from the account where the target database exists.

Multiple integrations limitations

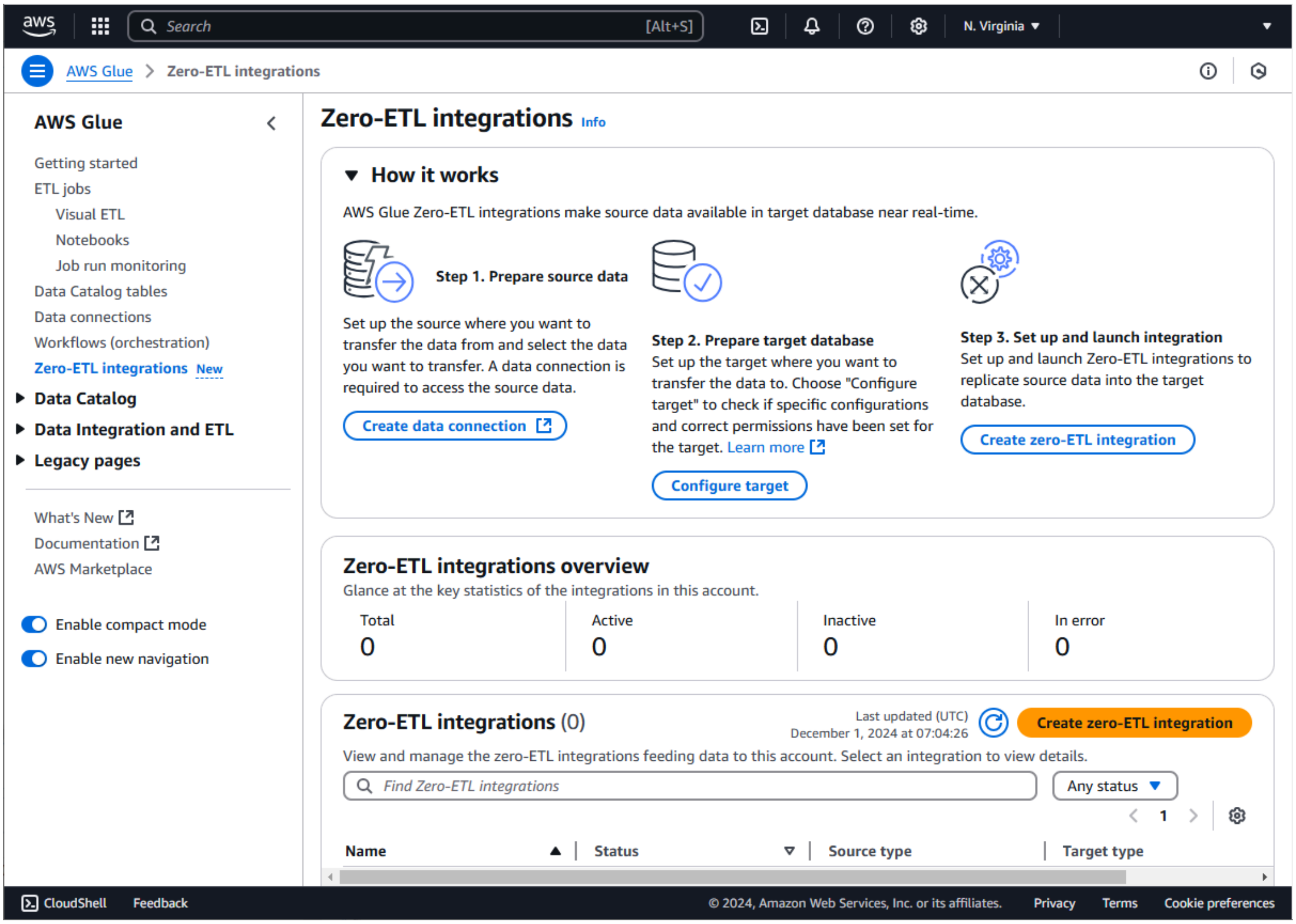
- If you need to replicate the same source with different schema unnest/partition configurations, it is required to create a new AWS Glue database for each integration separately. Later invoke CreateIntegrationTableProperty for each table from individual AWS Glue Database with desired schema unnesting and partitioning configurations.

Common integration tasks

Creating an integration

This sections describes the general steps to create an integration. This example uses Amazon DynamoDB as a source.

1. On the AWS Glue console home page, select **Zero-ETL integrations**.
2. You can view all you integrations on the Zero ETL integration home page. To create a new integration, select **Create zero-ETL integration**.


aws
Search [Alt+S]
N. Virginia ▼
AWS Glue > Zero-ETL integrations
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Zero-ETL integrations New
Data Catalog
Data Integration and ETL
Legacy pages
What's New
Documentation
AWS Marketplace
Enable compact mode
Enable new navigation
Zero-ETL integrations Info
How it works
Step 1. Prepare source data
Create data connection
Step 2. Prepare target database
Configure target
Step 3. Set up and launch integration
Create zero-ETL integration
Zero-ETL integrations overview
Total 0
Active 0
Inactive 0
In error 0
Zero-ETL integrations (0)
Create zero-ETL integration
Find Zero-ETL integrations
Any status ▼
1
Name
Status
Source type
Target type
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

AWS Glue Zero-ETL integrations make source data available in target database near real-time.

Set up the source where you want to transfer the data from and select the data you want to transfer. A data connection is required to access the source data.

Set up the target where you want to transfer the data to. Choose "Configure target" to check if specific configurations and correct permissions have been set for the target. Learn more

Set up and launch Zero-ETL integrations to replicate source data into the target database.

Glance at the key statistics of the integrations in this account.

Last updated (UTC) December 1, 2024 at 07:04:26

View and manage the zero-ETL integrations feeding data to this account. Select an integration to view details.

3. You are prompted to select a **Source Type**. Select your source and click **Next**. Refer to the the source configuration sections for SaaS integration sources.
4. In the **Configure source and target** page, select the tables or entities to replicate. For Amazon DynamoDB make sure the PITR and RBAC policy is configured.
5. Specify your integration target:
   - For an AWS Glue Data Catalog target, select the AWS Glue database you want to replicate the data to.

- For an Amazon Redshift data warehouse target, select the Redshift cluster namespace or Redshift Serverless workgroup namespace.

For more information, see Configuring the integration with your target.

6. Provide the **Target IAM Role** that you created in the prerequisites.
7. If you want to configure an optional **Target KMS Key** for your data being stored in the target, provide an enabled KMS Key. Likewise, if you want to configure a target network connection, select an AWS Glue connection.
8. The **Fix Target** button configures some of the steps in the Prerequisite section of this documentation. Namely it will 1) provide a Catalog RBAC Policy and 2) if no Amazon S3 URI is provided it will generate one for you, otherwise it will use the provided URI.
9. For integrations with a Redshift data warehouse target:
10. In the **Output setting** section of the **Configure source and target** page, select a schema unnesting option that you want for your data in the target. If you want to use customer partition keys for your data, select **Specify custom partition keys** and provide up to 10 keys. Otherwise, you can simply use the partition keys that are assigned to your DynamoDB table being replicated.
11. In the **Security and data encryption** section, you can provide a KMS key that will be used in the intermediary process of replicating your data to the target. Otherwise, an AWS managed KMS key will be used. Currently we only support a 15 minute replication setting. Enter a name for the Zero ETL integration in **Integration details**.
12. Review and make sure that all the provided details are correct. Click **Create and launch integration** once everything has been confirmed.
13. In the Zero ETL home page, you can select the integration you created and the details for your integrations will appear. The "Status" indicates the state of your integration.

Modifying an integration

You can modify an existing integration.

1. Select **Edit** in the top right corner of your integration details page.
2. On the **Edit source and target** page you can change the Target IAM role and Target network connection. The other fields are not editable after integration creation. Click **Next**.
3. You can also edit the name and description of the integration in the **Edit integration and configuration** page. Click **Next**.

4. Review your edits and once confirmed, click **Update integration**.

Deleting an integration

Delete is a terminal state for an integration. Once deleted, the integration cannot be revived. Deleting an integration clears up all internal metadata and any intermediate stored data.

During this process any running tasks which are writing data to a target table are terminated. AWS Glue will not delete or cleanup the target AWS Glue database (in the Data Catalog) and the associated data in the Amazon S3 bucket in your account. You need to explicitly clean those up if required.

To delete an integration:

1. In the integration details page, click **Delete**.
2. Enter "Delete" and click **Delete**. Note: This is an irreversible action.
3. In the integration details page, the status shows "Deleting". Once the integration is actually deleted, it will no longer appear on the the Zero ETL integration home page.

Creating integrations using the APIs

You can use the following APIs to create and manage Zero-ETL integrations in AWS Glue:

- CreateIntegration
- CreateIntegrationTableProperties
- CreateIntegrationResourceProperty
- UpdateIntegrationTableProperties
- UpdateIntegrationResourceProperty
- ModifyingIntegration
- DeleteIntegration
- DeleteIntegrationTableProperties
- DescribeIntegrations
- DescribeInboundIntegrations
- GetIntegrationTableProperties
- GetIntegrationResourceProperty

For more information, see Integration APIs in AWS Glue.

Monitoring an integration

Integration states

The following integration states describe the integration:

- `Creating` - The integration is being created.
- `Active` - The integration is sending transactional data to the target.
- `Modifying` - The integration is being modified.
- `Syncing` - The integration has encountered a recoverable error and is re-seeding data.
- `Needs attention` - The integration encountered an event or error that requires manual intervention to resolve it. To fix the issue, follow the instructions in the error message on the integration details.
- `Failed` - The integration encountered an unrecoverable event or error. You must delete and recreate the integration.
- `Deleting` - The integration is being deleted.

Viewing Amazon CloudWatch logs for an integration

AWS Glue zero-ETL integrations generate Amazon CloudWatch logs for visibility into your data movement. Log events relating to each successful ingestion or any failures experienced due to problematic data records at source, or data write errors due to schema changes or insufficient permissions are emitted to a default log group created in a customer account.

For each integration created, the log events for that integration will be collected under `/aws-glue/zeroETL-integrations/logs/` in Amazon Cloudwatch. Inside the log group, log messages will be split into log streams. Each integration created has a dedicated log stream to where all logs for that integration are written. For example, logs for an integration with IntegrationArn `arn:aws:glue:us-east-1:123456789012:integration:03cabe77-79e7-4b7a-b3da-8c160bea6bbf` can be found under /aws-glue/zeroETL-integrations/logs/03cabe77-79e7-4b7a-b3da-8c160bea6bbf. {IntegrationId} can be referenced from the {integrationArn} generated when an integration is created.

> **Note**
>
> For a cross-account scenario, source processing Logs are emitted in the source account where the integration exists and target processing logs are emitted in the target account where the target database exists.

IAM permissions required to enable logging

When creating your integration, the following IAM permissions are needed by the source and target roles to enable CloudWatch logging for an integration. AWS Glue zero-ETL integrations use these permissions provided in the source and target roles to emit CloudWatch logs to customer accounts.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "logs:CreateLogGroup",
        "logs:CreateLogStream",
        "logs:PutLogEvents"
      ],
      "Resource": [
        "*"
      ]
    }
  ]
}
```

Log messages

Log format: zero-ETL integrations emit four types of log messages:

```
// Ingestion started
{
"integrationArn": "arn:aws:glue:us-east-2:123456789012:integration/1a012bba-123a-1bba-
ab1c-173de3b12345",
...
    "messageType": "IngestionStarted",
    "details": {
        "tableName": "testDDBTable",
        "message": "Ingestion Job started"
    }
}
// Data processing stats on successful table ingestion
{
...
    "messageType": "IngestionProcessingStats",
    "details": {
        "tableName": "testDDBTable",
        "insert_count": 100,
        "update_count": 10,
        "delete_count": 10
    }
}
// Ingestion failure logs for failed table-processing
{
...
    "messageType": "IngestionFailed",
    "details": {
        "tableName": "testDDBTable",
        "errorMessage": "Failed to ingest data with error: Target Glue database not found.",
        "error_code" : "client_error"
    }
}
// Ingestion completed notification with lastSyncedTimestamp
{
...
    "messageType": "IngestionCompleted",
    "details": {
        "tableName": "testDDBTable",
        "message": "Ingestion Job completed"
        "lastSyncedTimestamp": "1132344255745"
    }
}
```

Viewing Amazon CloudWatch metrics for an integration

Once an integration completes, you can see these Amazon Cloudwatch metrics generated in your account for each AWS Glue job run:

CloudWatch metrics namespace: "AWS/Glue/ZeroETL"

Metrics dimensions:

- `integrationArn`
- `loadType`
- `tableName`

Metric names:

- `InsertCount` - number of records inserted in the target Iceberg table.
- `UpdateCount` - number of records updated in the target Iceberg table.
- `DeleteCount` - number of records deleted from the target Iceberg table.
- `IngestionSucceeded` - count 1, if the ingestion succeeded for the integration.
- `IngestionFailed` - count 1, if the ingestion failed for the integration.
- `LastSyncTimestamp` - timestamp until which source has been synced to target.

Managing event notifications with Amazon EventBridge

Zero-ETL integrations use Amazon EventBridge to manage event notifications to keep you up-to-date regarding changes in your integrations. Amazon EventBridge is a serverless event bus service that you can use to connect your applications with data from a variety of sources. In this case, the event source is AWS Glue. Events, which are monitored changes in an environment, are sent to EventBridge from AWS Glue automatically. Events are delivered in near real time.

EventBridge provides an environment for you to write event rules, which can specify actions to take for specific events. You can also set up targets, which are resources that EventBridge can send an event to. A target can include an API destination, an Amazon CloudWatch log group, and others. For more information about rules, see Amazon EventBridge rules. For more information about targets, see Amazon EventBridge targets.

To capture all Zero-ETL notifications, create an Eventbridge rule which matches the following:

```
{
  "source": [{
    "prefix": "aws.glue-zero-etl“
  }],
  "detail-type": [{
    "prefix": "Glue Zero ETL“
  }]
}
```

The following table includes zero-ETL integration events, with additional metadata:

Customer-facing detail type	Explanation
AWS Glue Zero ETL Ingestion Completed	Individual execution for an entity has completed successfully.
AWS Glue Zero ETL Ingestion Failed	Individual execution for an entity has completed unsuccessfully (either with a client or system error).
AWS Glue Zero ETL Integration Resynced	Integration has been RESYNCED.
AWS Glue Zero ETL Integration Failed	Integration status has changed to FAILED due to an error.
AWS Glue Zero ETL Integration Needs Attention	Integration status has changed to NEEDS_ATTENTION due to an error.
AWS Glue Zero ETL Ingestion In Progress	Individual execution for an entity has made partial progress towards completion.

Limitations

The following are general limitations of or considerations about zero-ETL integrations:

- Resource properties have a one-to-one relationship with the corresponding resource. Consequently, all integrations created using that resource must adhere to the singular resource property. Modifying a resource property will therefore impact all integrations associated with that resource.

- Table properties have a one-to-one relationship with the corresponding table or object within a resource. As a result, all integrations that process the same table to or from the same resource must adhere to the singular table property.
- You cannot rename a column at the source. If a column is renamed then there is no guarantee that the schema detection will be accurately done by AWS Glue and the repercussions on the integration are undefined.
- The following consideration applies to how the integration works with AWS Lake Formation managed tables: By default, you use the IAM/AWS Glue policy to manage your table and databases.
- If you want to use AWS Lake Formation to manage the table creation in that database, you need to make sure that the role is given sufficient Lake Formation permissions to create, modify and delete the table and database.
- The zero-ETL summary page does not contain any metrics at this time.

The following are source-specific limitations of zero-ETL integrations:

- Zero-ETL integrations with an SAP OData source now supports entities starting with `EntityOf`. The ability to override the primary key is currently supported only for SAPOData `EntityOf` objects. Once this property has been set, it cannot be modified.
- Zero-ETL integrations from Amazon DynamoDB to Amazon SageMaker Lakehouse (via S3) support a maximum DynamoDB table size of 50TB.
- The source DynamoDB table must be encrypted with either an Amazon-owned or customer-managed AWS KMS key. Amazon managed encryption is not supported for the source DynamoDB table.
- SAP OData works using a delta token, where the combination of an OAuth client plus an entity, or basic authentication plus an entity, can have only a single delta token. Avoid using the same entity in two different integrations with the same client.
- The following Salesforce entities or fields are unsupported for use in a zero-ETL integration with a Salesforce source.

```
AccountChangeEvent, AccountContactRoleChangeEvent, AccountHistory, AccountShare,
 ActiveFeatureLicenseMetric, ActivePermSetLicenseMetric, ActiveProfileMetric,
 ActivityFieldHistory, amzsec__asi_Telemetry_Data_Store__ChangeEvent,
 amzsec__asi_Telemetry_Data_Store__History, amzsec__asi_Telemetry_Data_Store__Share,
 amzsec__asi_Telemetry_Job_Log__ChangeEvent, amzsec__asi_Telemetry_Job_Log__History,
 amzsec__asi_Telemetry_Job_Log__Share,
amzsec__asi_Telemetry_Requirement__ChangeEvent,
amzsec__asi_Telemetry_Requirement__History,
amzsec__asi_Telemetry_Requirement__Share, ApexClass, ApexComponent, ApexLog,
ApexPage, ApexTestQueueItem, ApexTestResult, ApexTrigger, AssetChangeEvent,
AssetHistory, AssetRelationshipHistory, AssetShare, AssignmentRule,
AssociatedLocationHistory, AsyncApexJob, AuditTrailFileExportShare,
AuthorizationFormConsentChangeEvent, AuthorizationFormConsentHistory,
AuthorizationFormConsentShare, AuthorizationFormDataUseHistory,
AuthorizationFormDataUseShare, AuthorizationFormHistory, AuthorizationFormShare,
AuthorizationFormTextHistory, AuthProvider, AuthSession, BatchJobHistory,
BatchJobPartFailedRecordHistory, BatchJobPartHistory, BatchJobShare,
BrandTemplate, BriefcaseAssignmentChangeEvent, BriefcaseDefinitionChangeEvent,
BusinessBrandShare, BusinessHours, BusinessProcess, CalcMatrixColumnRangeHistory,
CalcProcStepRelationshipHistory, CalculationMatrixColumnHistory,
CalculationMatrixHistory, CalculationMatrixRowHistory, CalculationMatrixShare,
CalculationMatrixVersionHistory, CalculationProcedureHistory,
CalculationProcedureShare, CalculationProcedureStepHistory,
CalculationProcedureVariableHistory, CalculationProcedureVersionHistory, Calendar,
CalendarViewShare, CallCenter, CallCoachConfigModifyEvent, CampaignChangeEvent,
CampaignHistory, CampaignMemberChangeEvent, CampaignMemberStatusChangeEvent,
CampaignShare, CaseChangeEvent, CaseHistory, CaseHistory2 CaseHistory2ChangeEvent,
CaseRelatedIssueChangeEvent, CaseRelatedIssueHistory, CaseShare, CaseStatus,
CaseTeamMember, CaseTeamRole, CaseTeamTemplate, CaseTeamTemplateMember,
CaseTeamTemplateRecord, CategoryNode, ChangeRequestChangeEvent,
ChangeRequestHistory, ChangeRequestRelatedIssueChangeEvent,
ChangeRequestRelatedIssueHistory, ChangeRequestRelatedItemChangeEvent,
ChangeRequestRelatedItemHistory, ChangeRequestShare, ChatRetirementRdyMetrics,
ChatterActivity, ClientBrowser, CollaborationGroup, CollaborationGroupMember,
CollaborationGroupMemberRequest, CollaborationInvitation,
CommSubscriptionChannelTypeHistory, CommSubscriptionChannelTypeShare,
CommSubscriptionConsentChangeEvent, CommSubscriptionConsentHistory,
CommSubscriptionConsentShare, CommSubscriptionHistory, CommSubscriptionShare,
CommSubscriptionTimingHistory, Community, ConnectedApplication, ContactChangeEvent,
ContactHistory, ContactPointAddressChangeEvent, ContactPointAddressHistory,
ContactPointAddressShare, ContactPointConsentChangeEvent,
ContactPointConsentHistory, ContactPointConsentShare,
ContactPointEmailChangeEvent, ContactPointEmailHistory, ContactPointEmailShare,
ContactPointPhoneChangeEvent, ContactPointPhoneHistory, ContactPointPhoneShare,
ContactPointTypeConsentChangeEvent, ContactPointTypeConsentHistory,
ContactPointTypeConsentShare, ContactRequestShare, ContactShare,
ContentDocumentChangeEvent, ContentDocumentHistory, ContentDocumentLink,
ContentDocumentLinkChangeEvent, ContentDocumentSubscription, ContentFolderItem,
ContentFolderLink, ContentFolderMember, ContentNote, ContentNotification,
ContentTagSubscription, ContentUserSubscription, ContentVersionChangeEvent,
ContentVersionComment, ContentVersionHistory, ContentVersionRating,
ContentWorkspace, ContentWorkspaceMember, ContentWorkspacePermission,
ContentWorkspaceSubscription, ContractChangeEvent, ContractHistory,
ContractLineItemChangeEvent, ContractLineItemHistory, ContractStatus,
Conversation, ConversationParticipant, CronJobDetail, CronTrigger, CustomBrand,
CustomBrandAsset, CustomerShare, CustomHttpHeader, DashboardComponent,
DataUseLegalBasisHistory, DataUseLegalBasisShare, DataUsePurposeHistory,
DataUsePurposeShare, DecisionTableRecordset, DeleteEvent, DocumentAttachmentMap,
Domain, DomainSite, DTRecordsetReplicaShare, EmailBounceEvent,
EmailMessageChangeEvent, EmailServicesAddress, EmailServicesFunction,
EmailTemplate, EmailTemplateChangeEvent, EngagementAttendeeChangeEvent,
EngagementAttendeeHistory, EngagementChannelTypeHistory, EngagementChannelTypeShare,
EngagementInteractionChangeEvent, EngagementInteractionHistory,
EngagementInteractionShare, EngagementInterface, EngagementTopicChangeEvent,
EngagementTopicHistory, EntitlementChangeEvent, EntitlementHistory,
EntitlementTemplate, EntityMilestoneHistory, EntitySubscription, EventChangeEvent,
EventRelationChangeEvent, EventRelayConfigChangeEvent, ExpressionSetHistory,
ExpressionSetShare, ExpressionSetVersionHistory, ExternalEventMappingShare,
FeedAttachment, FeedLike, FeedPollChoice, FeedPollVote, FeedSignal,
FieldPermissions, FieldSecurityClassification, FiscalYearSettings,
FlowInterviewLogShare, FlowInterviewShare, FlowOrchestrationEvent,
FlowOrchestrationInstanceShare, FlowOrchestrationStageInstanceShare,
FlowOrchestrationStepInstanceShare, FlowOrchestrationWorkItemShare,
FlowRecordShare, FlowRecordVersionChangeEvent, FlowTestResultShare, Folder,
Group, GroupMember, Holiday, IdeaComment, IdpEventLog, ImageHistory, ImageShare,
IncidentChangeEvent, IncidentHistory, IncidentRelatedItemChangeEvent,
IncidentRelatedItemHistory, IncidentShare, IndividualChangeEvent,
IndividualHistory, IndividualShare, KnowledgeableUser, LeadChangeEvent, LeadHistory,
LeadShare, LeadStatus, LightningExitByPageMetrics, LightningToggleMetrics,
LightningUsageByAppTypeMetrics, LightningUsageByBrowserMetrics,
LightningUsageByFlexiPageMetrics, LightningUsageByPageMetrics, ListEmailChangeEvent,
ListEmailShare, ListView, LocationChangeEvent, LocationHistory, LocationShare,
LocationTrustMeasureShare, LoginHistory, LoginIp, MacroChangeEvent,
MacroHistory, MacroInstructionChangeEvent, MacroShare, MacroUsageShare,
ManagedContentVariantChangeEvent, MessagingEndUserHistory, MessagingEndUserShare,
MessagingSessionHistory, MessagingSessionShare, MilestoneType, MLEngagementEvent,
ObjectPermissions, OpportunityChangeEvent, OpportunityContactRoleChangeEvent,
OpportunityFieldHistory, OpportunityLineItemChangeEvent, OpportunityShare,
OpportunityStage, OrderChangeEvent, OrderHistory, OrderItemChangeEvent,
OrderItemHistory, OrderShare, OrderStatus, Organization, OrgEmailAddressSecurity,
OrgWideEmailAddress, OutgoingEmail, OutgoingEmailRelation,
PackageLicense, PartnerRole, PartyConsentChangeEvent, PartyConsentHistory,
PartyConsentShare, Period, PermissionSet, PermissionSetAssignment,
PermissionSetTabSetting, Pricebook2ChangeEvent, Pricebook2History,
 PricebookEntryChangeEvent, PricebookEntryHistory, PrivacyJobSessionShare,
 PrivacyObjectSessionShare, PrivacyRTBFRequestHistory, PrivacyRTBFRequestShare,
 PrivacySessionRecordFailureShare, ProblemChangeEvent, ProblemHistory,
 ProblemIncidentChangeEvent, ProblemIncidentHistory, ProblemRelatedItemChangeEvent,
 ProblemRelatedItemHistory, ProblemShare, ProcessDefinition, ProcessExceptionEvent,
 ProcessExceptionShare, ProcessInstanceChangeEvent, ProcessInstanceStep,
 ProcessInstanceStepChangeEvent, ProcessNode, Product2ChangeEvent, Product2History,
 ProductEntitlementTemplate, Profile, ProfileSkillEndorsementHistory,
 ProfileSkillHistory, ProfileSkillShare, ProfileSkillUserHistory, PromptActionShare,
 PromptErrorShare, QueueSobject, QuickTextChangeEvent, QuickTextHistory,
 QuickTextShare, QuickTextUsageShare, RecentlyViewed, RecommendationChangeEvent,
 RecordActionHistory, RecordAlertHistory, RecordAlertShare, RecordType,
 ScorecardShare, SellerHistory, SellerShare, ServiceContractChangeEvent,
 ServiceContractHistory, ServiceContractShare, SetupAuditTrail, SetupEntityAccess,
 SharingRecordCollectionShare, Site, SiteHistory, SiteRedirectMapping,
 SocialPersonaHistory, SocialPostChangeEvent, SocialPostHistory, SocialPostShare,
 SolutionHistory, SolutionStatus, StaticResource, StreamingChannelShare,
 TableauHostMappingShare, TaskChangeEvent, TaskPriority, TaskStatus,
 ThreatDetectionFeedback, TimelineObjectDefinitionChangeEvent, TodayGoalShare,
 Topic, TopicUserEvent, Translation, User, UserAppMenuCustomizationShare,
 UserChangeEvent, UserDefinedLabelAssignmentShare, UserDefinedLabelShare,
 UserEmailPreferredPersonShare, UserLicense, UserLogin, UserPackageLicense,
 UserPreference, UserPrioritizedRecordShare, UserProvisioningRequestShare,
 UserRole, UserShare, VideoCallChangeEvent, VideoCallParticipantChangeEvent,
 VideoCallRecordingChangeEvent, VideoCallShare, VisualforceAccessMetrics,
 VoiceCallChangeEvent, VoiceCallRecordingChangeEvent, VoiceCallShare, Vote,
 WebLink, WorkAccessShare, WorkBadgeDefinitionHistory, WorkBadgeDefinitionShare,
 WorkOrderChangeEvent, WorkOrderHistory, WorkOrderLineItemChangeEvent,
 WorkOrderLineItemHistory, WorkOrderLineItemStatus, WorkOrderShare,
 WorkOrderStatus, WorkPlanChangeEvent, WorkPlanHistory, WorkPlanShare,
 WorkPlanTemplateChangeEvent, WorkPlanTemplateEntryChangeEvent,
 WorkPlanTemplateEntryHistory, WorkPlanTemplateHistory, WorkPlanTemplateShare,
 WorkStepChangeEvent, WorkStepHistory, WorkStepStatus, WorkStepTemplateChangeEvent,
 WorkStepTemplateHistory, WorkStepTemplateShare, WorkThanksShare
```

- The following ServiceNow entities or fields are unsupported for use in a zero-ETL integration with a ServiceNow source.

```
ais_acl_overrides, ais_async_genius_result, ais_async_request,
 ais_connection, ais_genius_result_configuration_parameters,
 ais_partition_health, ais_partition_health_response, ais_publish_history,
 ais_relevancy_training_execution, ais_relevancy_training_staging,
 ais_search_profile_relevancy_model, catalog_draft_entities, clone_log,
 clone_log0000, clone_log0001, clone_log0002, clone_log0003, clone_log0004,
 clone_log0005, clone_log0006, clone_log0007, cmdb_ie_context, cmdb_ie_log,
 cmdb_ie_run, cmdb_ire_partial_payloads_index, cmdb_qb_result_base
$par1, cmdb$par1, discovery_log, discovery_log0000, discovery_log0001,
 discovery_log0002, discovery_log0003, discovery_log0004, discovery_log0005,
 discovery_log0006, discovery_log0007, ecc_agent_log, entitlement_data,
 entl_subscription_map, gs_entitlement_plugin_mapping, ih_transaction_exclusion,
 import_log, import_log0000, import_log0001, import_log0002, import_log0003,
 import_log0004, import_log0005, import_log0006, import_log0007, jrobin_archive,
 jrobin_database, jrobin_datasource, jrobin_definition, jrobin_graph,
 jrobin_graph_line, jrobin_graph_set, jrobin_graph_set_member, jrobin_shard,
 jrobin_shard_location, license_role_discovery_run, logger_configuration_validation,
 m2m_analytics_event_logger, m2m_user_consent_info, ml_artifact_object_store,
 np$sys_gen_ai_filter_sample, np$sys_ui_element, np$sys_ui_list_element,
 one_api_service_plan_feature_invocation, one_api_service_plan_invocation,
 open_nlu_predict_log, open_nlu_predict_log0000, open_nlu_predict_log0001,
 open_nlu_predict_log0002, open_nlu_predict_log0003, open_nlu_predict_log0004,
 open_nlu_predict_log0005, open_nlu_predict_log0006, open_nlu_predict_log0007,
 pa_diagnostic_log, pa_diagnostic_log0000, pa_diagnostic_log0001,
 pa_diagnostic_log0002, pa_diagnostic_log0003, pa_diagnostic_log0004,
 pa_diagnostic_log0005, pa_diagnostic_log0006, pa_diagnostic_log0007, pa_favorites,
 pa_job_log_rows, pa_job_log_rows0000, pa_job_log_rows0001, pa_job_log_rows0002,
 pa_job_log_rows0003, pa_job_log_rows0004, pa_job_log_rows0005, pa_job_log_rows0006,
 pa_job_log_rows0007, pa_migration_ignored_scores, pa_scores_l1, pa_scores_l2,
 pa_scores_migration_groups, par_dashboard_conversion_backup, promin_log,
 promin_log0000, promin_log0001, promin_log0002, promin_log0003, promin_log0004,
 promin_log0005, promin_log0006, promin_log0007, promin_request_object,
 proposed_change_verification_log, proposed_change_verification_log0000,
 proposed_change_verification_log0001, proposed_change_verification_log0002,
 proposed_change_verification_log0003, proposed_change_verification_log0004,
 proposed_change_verification_log0005, proposed_change_verification_log0006,
 proposed_change_verification_log0007, protected_table_log, protected_table_log0000,
 protected_table_log0001, protected_table_log0002, protected_table_log0003,
 protected_table_log0004, protected_table_log0005, protected_table_log0006,
 protected_table_log0007, pwd_history, qb_query_results, scan_log,
 scan_log0000, scan_log0001, scan_log0002, scan_log0003, scan_log0004,
 scan_log0005, scan_log0006, scan_log0007, schema_validator_error,
 sla_repair_log_entry, sla_repair_log_entry0000, sla_repair_log_entry0001,
 sla_repair_log_entry0002, sla_repair_log_entry0003, sla_repair_log_entry0004,
 sla_repair_log_entry0005, sla_repair_log_entry0006, sla_repair_log_entry0007,
 sla_repair_log_message, sla_repair_log_message0000, sla_repair_log_message0001,
 sla_repair_log_message0002, sla_repair_log_message0003, sla_repair_log_message0004,
 sla_repair_log_message0005, sla_repair_log_message0006, sla_repair_log_message0007,
 sn_bm_client_activity, sn_ci_analytics_st_actionable_notifs,
sn_ci_analytics_st_conv_completion_by_cat, sn_ci_analytics_st_conv_dynamic_property,
sn_ci_analytics_st_conversation, sn_ci_analytics_st_count_by_date,
sn_ci_analytics_st_event_occurrence, sn_ci_analytics_st_event_property_value_trend,
sn_ci_analytics_st_issue_auto_resolution, sn_ci_analytics_st_no_clicks,
sn_ci_analytics_st_no_results, sn_ci_analytics_st_property_summary_by_event,
sn_ci_analytics_st_session_count_per_locale, sn_ci_analytics_st_session_duration,
sn_ci_analytics_st_spokes_usage, sn_ci_analytics_st_topic_execution_stats,
sn_ci_analytics_st_topic_occurrence, sn_ci_analytics_st_trending_content,
sn_ci_analytics_st_trending_queries, sn_ci_analytics_st_users,
sn_cs_plugin_signatures, sn_cs_telemetry_log, sn_dfc_application, sn_dfc_product,
sn_employee_position, sn_entitlement_st_subscription_application_family,
sn_entitlement_st_subscription_application_users, sn_hr_sp_st_relevant_for_you,
sn_instance_clone_log, sn_instance_clone_log0000, sn_instance_clone_log0001,
sn_instance_clone_log0002, sn_instance_clone_log0003, sn_instance_clone_log0004,
sn_instance_clone_log0005, sn_instance_clone_log0006, sn_instance_clone_log0007,
sn_km_mr_st_kb_knowledge, sn_me_st_topic, sn_rf_conditional_definition,
sn_rf_evaluation_type, sn_rf_evaluation_type_input, sn_rf_recommendation_action,
sn_rf_recommendation_experience, sn_rf_recommendation_history,
sn_rf_recommendation_rule, sn_rf_record_display_configuration,
sn_rf_trend_definition, sn_sub_man_st_account_level_entitlement,
sn_sub_man_st_gen_ai_metadata, sn_sub_man_st_instance_used_assist_count,
sn_sub_man_st_now_assist_creator_instances, sn_sub_man_st_now_assists_aggregate,
sn_sub_man_st_subscribed_groups, sn_sub_man_st_subscription_insights,
sn_sub_man_st_subscription_license_detail_metric,
sn_sub_man_st_unallocated_group_recommendation,
sn_sub_man_st_unconfirmed_user_group, sn_wn_user_app_activity,
sn_wn_user_content_activity, snc_monitorable_item,
snpar_sched_export_v_scheduled_export_visualization, spotlight,
spotlight_audit, spotlight_copy_log_row, spotlight_copy_log_row0000,
spotlight_copy_log_row0001, spotlight_copy_log_row0002, spotlight_copy_log_row0003,
spotlight_copy_log_row0004, spotlight_copy_log_row0005, spotlight_copy_log_row0006,
spotlight_copy_log_row0007, spotlight_job_log_row, spotlight_job_log_row0000,
spotlight_job_log_row0001, spotlight_job_log_row0002, spotlight_job_log_row0003,
spotlight_job_log_row0004, spotlight_job_log_row0005, spotlight_job_log_row0006,
spotlight_job_log_row0007, st_dfc_performance_metric, st_license_detail_metric,
st_on_call_hour, st_sc_wizard_question, st_sys_catalog_items_and_variable_sets,
st_sys_design_system_icon, subscription_instance_stats,
svc_container_config, svc_environment_config, svc_layer_config,
svc_model_assoc_ci, svc_model_checkpoint_attr, svc_model_obj_cluster,
svc_model_obj_constraint, svc_model_obj_deployable, svc_model_obj_element,
svc_model_obj_impact, svc_model_obj_impactrule, svc_model_obj_package,
svc_model_obj_path, svc_model_obj_relation, svc_model_obj_service,
sys_administrative_script_transaction, sys_amb_message, sys_amb_processor,
sys_analytics_batch_state, sys_analytics_config, sys_analytics_data_points_error,
sys_analytics_event, sys_analytics_logger, sys_analytics_logger_field,
sys_app_payload_loader_rule, sys_app_payload_unloader_rule, sys_app_scan_payload,
sys_app_scan_variable, sys_app_scan_variable_type, sys_archive_destroy_log,
sys_archive_destroy_run, sys_archive_log, sys_archive_run, sys_atf_transaction_log,
sys_attachment_doc, sys_attachment_doc_v2, sys_attachment_soft_deleted, sys_audit,
sys_audit_relation, sys_auth_policy_api_allowed, sys_aw_registered_scripting_modal,
sys_cache_flush, sys_data_egress_source, sys_dm_delete_count,
sys_export_set_log, sys_export_set_log0000, sys_export_set_log0001,
sys_export_set_log0002, sys_export_set_log0003, sys_export_set_log0004,
sys_export_set_log0005, sys_export_set_log0006, sys_export_set_log0007,
sys_flow_compiled_flow, sys_flow_compiled_flow_chunk, sys_flow_context_chunk,
sys_flow_context_chunk_archive, sys_flow_context_inputs_chunk,
sys_flow_execution_history, sys_flow_log, sys_flow_log0000, sys_flow_log0001,
sys_flow_log0002, sys_flow_log0003, sys_flow_plan_context_binding,
sys_flow_report_doc, sys_flow_report_doc_chunk, sys_flow_report_doc_chunk_archive,
sys_flow_runtime_state_chunk, sys_flow_runtime_value_chunk,
sys_flow_subflow_plan_chunk, sys_flow_trigger_plan_chunk, sys_flow_val_listener,
sys_flow_value, sys_flow_value_chunk, sys_gen_ai_config_example,
sys_gen_ai_feature_mapping, sys_gen_ai_strategy_mapping, sys_gen_ai_usage_log,
sys_generative_ai_capability_definition, sys_generative_ai_log,
sys_generative_ai_response_validator, sys_generative_ai_validator,
sys_geo_routing, sys_geo_routing_config, sys_hop_token, sys_hub_action_plan_chunk,
sys_hub_snapshot_chunk, sys_journal_field, sys_journal_field_edit, sys_json_chunk,
sys_kaa_policy, sys_kaa_subidentity_assertion, sys_kaa_user_policy_mapping,
sys_mapplication, sys_mass_encryption_job, sys_notification_execution_log,
sys_notification_execution_log0000, sys_notification_execution_log0001,
sys_notification_execution_log0002, sys_notification_execution_log0003,
sys_notification_execution_log0004, sys_notification_execution_log0005,
sys_notification_execution_log0006, sys_notification_execution_log0007,
sys_nowmq_message, sys_nowmq_provider_param_definition, sys_orchestrator_action,
sys_pd_asset_configuration, sys_pd_context_chunk, sys_pd_context_log,
sys_pd_snapshot_chunk, sys_pd_trigger_license, sys_processing_framework_job,
sys_query_index_hint, sys_query_rewrite, sys_query_string_log,
sys_replication_queue, sys_replication_queue0, sys_replication_queue1,
sys_replication_queue2, sys_replication_queue3, sys_replication_queue4,
sys_replication_queue5, sys_replication_queue6, sys_replication_queue7,
sys_request_performance, sys_rollback_blacklisted, sys_rollback_conflict,
sys_rollback_incremental, sys_rollback_log, sys_rollback_log0000,
sys_rollback_log0001, sys_rollback_log0002, sys_rollback_log0003,
sys_rollback_log0004, sys_rollback_log0005, sys_rollback_log0006,
sys_rollback_log0007, sys_rollback_run, sys_rollback_schema_change,
sys_rollback_schema_conflict, sys_rollback_sequence, sys_scheduler_assignment,
sys_scheduler_memory_pressure_job_log, sys_script_adapter_rule,
sys_script_batch_adapter_rule, sys_search_source_filter, sys_service_authentication,
sys_signing_job, sys_suggestion_reader, sys_sync_history_review, sys_trend,
sys_unreferenced_preview, sys_unreferenced_record_rule, sys_upgrade_manifest,
sys_upgrade_state, sys_ux_asset_cache_buster, sys_ux_lib_component_prop,
sys_ux_lib_presource, sys_ux_page_action, sys_ux_page_action_binding,
sysevent_queue_runtime, syslog, syslog_app_scope0000, syslog_app_scope0001,
syslog_app_scope0002, syslog_app_scope0003, syslog_app_scope0004,
syslog_app_scope0005, syslog_app_scope0006, syslog_app_scope0007, syslog_email0000,
syslog_email0001, syslog_email0002, syslog_email0003, syslog_email0004,
syslog_email0005, syslog_email0006, syslog_email0007, syslog_transaction,
syslog_transaction0000, syslog_transaction0001, syslog_transaction0002,
syslog_transaction0003, syslog_transaction0004, syslog_transaction0005,
syslog_transaction0006, syslog_transaction0007, syslog0000, syslog0001,
syslog0002, syslog0003, syslog0004, syslog0005, syslog0006, syslog0007,
ts_attachment, ts_c_1_0, ts_c_1_1, ts_c_1_2, ts_c_1_3, ts_c_1_4, ts_c_1_5,
ts_c_1_6, ts_c_1_7, ts_c_1_8, ts_c_1_9, ts_c_10_0, ts_c_10_1, ts_c_10_2,
ts_c_10_3, ts_c_10_4, ts_c_10_5, ts_c_10_6, ts_c_10_7, ts_c_10_8, ts_c_10_9,
ts_c_11_0, ts_c_11_1, ts_c_11_2, ts_c_11_3, ts_c_11_4, ts_c_11_5, ts_c_11_6,
ts_c_11_7, ts_c_11_8, ts_c_11_9, ts_c_2_0, ts_c_2_1, ts_c_2_2, ts_c_2_3,
ts_c_2_4, ts_c_2_5, ts_c_2_6, ts_c_2_7, ts_c_2_8, ts_c_2_9, ts_c_3_0,
ts_c_3_1, ts_c_3_2, ts_c_3_3, ts_c_3_4, ts_c_3_5, ts_c_3_6, ts_c_3_7, ts_c_3_8,
ts_c_3_9, ts_c_39_0, ts_c_39_1, ts_c_39_2, ts_c_39_3, ts_c_39_4, ts_c_39_5,
ts_c_39_6, ts_c_39_7, ts_c_39_8, ts_c_39_9, ts_c_40_0, ts_c_40_1, ts_c_40_2,
ts_c_40_3, ts_c_40_4, ts_c_40_5, ts_c_40_6, ts_c_40_7, ts_c_40_8, ts_c_40_9,
ts_c_41_0, ts_c_41_1, ts_c_41_2, ts_c_41_3, ts_c_41_4, ts_c_41_5, ts_c_41_6,
ts_c_41_7, ts_c_41_8, ts_c_41_9, ts_c_42_0, ts_c_42_1, ts_c_42_2, ts_c_42_3,
ts_c_42_4, ts_c_42_5, ts_c_42_6, ts_c_42_7, ts_c_42_8, ts_c_42_9, ts_c_43_0,
ts_c_43_1, ts_c_43_2, ts_c_43_3, ts_c_43_4, ts_c_43_5, ts_c_43_6, ts_c_43_7,
ts_c_43_8, ts_c_43_9, ts_c_44_0, ts_c_44_1, ts_c_44_2, ts_c_44_3, ts_c_44_4,
ts_c_44_5, ts_c_44_6, ts_c_44_7, ts_c_44_8, ts_c_44_9, ts_c_45_0, ts_c_45_1,
ts_c_45_2, ts_c_45_3, ts_c_45_4, ts_c_45_5, ts_c_45_6, ts_c_45_7, ts_c_45_8,
ts_c_45_9, ts_c_6_0, ts_c_6_1, ts_c_6_2, ts_c_6_3, ts_c_6_4, ts_c_6_5,
ts_c_6_6, ts_c_6_7, ts_c_6_8, ts_c_6_9, ts_c_7_0, ts_c_7_1, ts_c_7_2, ts_c_7_3,
ts_c_7_4, ts_c_7_5, ts_c_7_6, ts_c_7_7, ts_c_7_8, ts_c_7_9, ts_c_8_0, ts_c_8_1,
ts_c_8_2, ts_c_8_3, ts_c_8_4, ts_c_8_5, ts_c_8_6, ts_c_8_7, ts_c_8_8, ts_c_8_9,
ts_c_attachment, ts_chain, ts_deleted_doc, ts_document, ts_field, ts_index_stats,
ts_phrase, ts_search_stats, ts_v4_attachment, ts_word, ua_app_metadata,
ua_audit_stats, ua_extra_page, ua_monitor_property, ua_monitor_property_audit,
ua_shared_service, ua_sn_table_inventory, ua_sp_known_bot, ua_upload_log,
v_ais_result_improvement_rule_condition_builder_values, v_cluster_nodes,
v_cxs_search_resource, v_db_index, v_db_trigger, v_file_load_order,
v_interaction_context, v_iostats, v_mysql_proclist, v_mysql_status,
v_mysql_variables, v_on_call_report_cache, v_pa_par_combined_dashboard,
v_pd_activity_condition_to_run, v_pd_activity_start_rule_with_condition,
v_pd_lane_condition_to_run, v_sql_debug, v_st_kb_category, v_st_kb_most_viewed,
v_st_kb_recently_viewed, v_st_km_genai_mra_similar_task, v_st_popular_item,
v_st_recent_item, v_st_sc_cat_item, v_st_sc_catalog, v_st_sc_category,
validator_run_summary, vtb_card_history, wf_log, wf_log0000, wf_log0001,
wf_log0002, wf_log0003, wf_log0004, wf_log0005, wf_log0006, wf_log0007,
spotlight_criteria, pa_snapshots, pa_widget_indicators, pa_widgets,
sn_cim_register, global, gsw_change_log, gsw_content, gsw_content_group,
gsw_content_information, gsw_status_of_content, multi_factor_browser_fingerprint,
multi_factor_criteria, pa_filters, password_policy, plan_execution,
plan_mysql, plan_oracle, plan_postgres, sc_rest_api_without_access_policy,
sn_actsub_activity, sn_actsub_activity_fanout, sn_actsub_activity_stream,
sn_actsub_activity_type, sn_actsub_atype_attributes, sn_actsub_atype_notif_pref,
sn_actsub_module, sn_actsub_notif_object, sn_actsub_subobject_stream,
sn_actsub_subscribable_object, sn_actsub_subscription_notif_pref,
sn_actsub_user_stream, sn_appclient_store_outbound_http_quota,
sn_appcreator_app_template, sn_critical_update, sn_docker_spoke_images,
sn_employee_app, sn_employee_app_access, sn_employee_app_access_criteria,
sn_entitlement_genai_assist_counts, sn_entitlement_genai_creator_user_counts,
sn_entitlement_genai_creator_users, sn_mif_instance, sn_mif_sync_data,
sn_mif_sync_status, sn_mif_table_registration, sn_mif_trust_config,
sn_vsc_best_practice_configurations, sn_vsc_best_practice_goals,
sn_vsc_changed_hardening_settings, sn_vsc_changed_scan_findings,
sn_vsc_check_security_area, sn_vsc_elevation_event, sn_vsc_event,
sn_vsc_export_event, sn_vsc_export_setting, sn_vsc_harc_compliance_status_lookup,
sn_vsc_hardening_compliance_scores, sn_vsc_impersonation_event,
sn_vsc_instance_hardening_settings, sn_vsc_login_event,
sn_vsc_scan_comparisons, sn_vsc_scan_summary, sn_vsc_security_check_categories,
sn_vsc_security_check_configurations, sn_vsc_security_configuration_groups,
sn_vsc_security_privacy_capabilities, sn_vsc_updated_settings,
sn_vsc_user_comparisons, sys_app_hash_inventory, sys_coalesce_strategy_deferred,
sys_flow_secure_data, sys_formula_function, sys_geocoding_request,
sys_global_file_hash, sys_import_set_row, sys_index, sys_index_explain,
sys_installation_schedule, sys_installation_schedule_item, sys_offline_app,
sys_package, sys_package_dependency_item, sys_package_dependency_m2m,
sys_plugins, sys_querystat, sys_reap_package, sys_scoped_plugin,
sys_stage_storage_alias, sys_storage_alias, sys_storage_table_alias, sys_store_app,
sys_table_partition, sys_upgrade_history_log, sys_user_public_credential,
sys_webauthn_authentication_request, sys_webauthn_registration_request,
syslog_app_scope, syslog_email, syslog_page_timing, ua_instance_state_config,
v_expression_cache, v_private_cache, v_shared_cache, sys_amb_message0002,
sys_amb_message0004, sys_amb_message0005, sys_metadata, v_par_unified_report_viz
```

AWS Glue Data Quality

AWS Glue Data Quality allows you to measure and monitor the quality of your data so that you can make good business decisions. Built on top of the open-source DeeQu framework, AWS Glue Data Quality provides a managed, serverless experience. AWS Glue Data Quality works with Data Quality Definition Language (DQDL), which is a domain specific language that you use to define data quality rules. To learn more about DQDL and supported rule types, see Data Quality Definition Language (DQDL) reference.

For additional product details and pricing, see the service page for AWS Glue Data Quality.

Benefits and key features

Benefits and key features of AWS Glue Data Quality include:

- **Serverless** – There is no installation, patching or maintenance.
- **Get started quickly** – AWS Glue Data Quality quickly analyzes your data and creates data quality rules for you. You can get started with two clicks: “Create Data Quality Rules → Recommend rules”.
- **Detect data quality issues** – Use machine learning (ML) to detect anomalies and hard-to-detect data quality issues.
- **Improvise your rules** – with 25+ out-of-the-box DQ rules to start from, you can create rules that suit your specific needs.
- **Evaluate quality and make confident business decisions** – Once you evaluate the rules, you get a Data Quality score that provides an overview of the health of your data. Use Data Quality score to make confident business decisions.
- **Zero in on bad data** – AWS Glue Data Quality helps you identify the exact records that caused your quality scores to go down. Easily identify them, quarantine and fix them.
- **Pay as you go** – There are no annual licenses you need to use AWS Glue Data Quality.
- **No lock-in** – AWS Glue Data Quality is built on open source DeeQu, allowing you to keep the rules you are authoring in an open language.
- **Data quality checks** – You can enforce data quality checks on Data Catalog and AWS Glue ETL pipelines allowing you to manage data quality at rest and in transit.
- **ML-based data quality detection** – Use machine learning (ML) to detect anomalies and hard-to-detect data quality issues.

- **Open language to express rules** – ensures that data quality rules are authored consistently and simply. Business users can easily express data quality rules in a straightforward language that they can understand. For engineers, this language provides the flexibility to generate code, implement consistent version control, and automate deployments.

How it works

There are two entry points for AWS Glue Data Quality: the AWS Glue Data Catalog and AWS Glue ETL jobs. This section provides an overview of the use cases and AWS Glue features that each entry point supports.

Data quality for the AWS Glue Data Catalog

AWS Glue Data Quality evaluates objects that are stored in the AWS Glue Data Catalog It offers non-coders an easy way to set up data quality rules. These personas include data stewards and business analysts.

You might choose this option for the following use cases:

- You want to perform data quality tasks on data sets that you've already cataloged in the AWS Glue Data Catalog.
- You work on data governance and need to identify or evaluate data quality issues in your data lake on an ongoing basis.

You can manage data quality for the Data Catalog using the following interfaces:

- The AWS Glue management console
- AWS Glue APIs

To get started with AWS Glue Data Quality for the AWS Glue Data Catalog see Getting started with AWS Glue Data Quality for the Data Catalog.

Data quality for AWS Glue ETL jobs

AWS Glue Data Quality for AWS Glue ETL jobs lets you perform *proactive* data quality tasks. Proactive tasks help you identify and filter out bad data *before* you load a data set into your data lake.

Video: Introducing AWS Glue Data Quality for ETL Pipelines

You might choose data quality for ETL jobs for the following use cases:

- You want to incorporate data quality tasks into your ETL jobs
- You want to write code that defines data quality tasks in ETL scripts
- You want to manage the quality of data that flows in your visual data pipelines

You can manage data quality for ETL jobs using the following interfaces:

- AWS Glue Studio, AWS Glue Studio notebooks, and AWS Glue interactive sessions
- AWS Glue libraries for ETL scripting
- AWS Glue APIs

To get started with data quality for ETL jobs, see Tutorial: Getting started with Data Quality in the *AWS Glue Studio User Guide*.

Comparing data quality for the Data Catalog to data quality for ETL jobs

This table provides an overview of features that each entry point for AWS Glue Data Quality supports.

Feature	Data quality for the Data Catalog	Data quality for ETL jobs
Data sources	Amazon S3, Amazon Redshift, JDBC sources compatible with the Data Catalog, and transactional data lake formats such as Apache	All data sources supported by AWS Glue, including custom connectors and third-party connectors.

Feature	Data quality for the Data Catalog	Data quality for ETL jobs
	Iceberg, Apache Hudi, and Delta Lake. AWS Lake Formation managed OTF formats are also supported with some limitations. Amazon Athena views that are cataloged in AWS Glue Data Catalog are not supported. Please see Supported source types.	
Data Quality rule recommendations	Supported	Not supported
Author and run DQDL rules	Supported	Supported
Auto scaling	Not supported	Supported
AWS Glue Flex support	Not supported	Supported
Scheduling	Supported when evaluating Data Quality rules and via Step Functions.	Supported when using Step Functions and workflows.
Identifying records that failed data quality checks	Not supported	Supported
Integration with Amazon Eventbridge	Supported	Supported
Integration with AWS Cloudwatch	Supported	Supported
Writing data quality results to Amazon S3	Supported	Supported

Feature	Data quality for the Data Catalog	Data quality for ETL jobs
Incremental data quality	Supported via pushdown predicates	Supported via AWS Glue bookmarks
AWS CloudFormation support	Supported	Supported
ML-based anomaly detection	Not supported	Supported
Dynamic rules	Not supported	Supported

Considerations

Consider the following items before you use AWS Glue Data Quality:

- Data quality rules can't evaluate nested or list-type data sources. See Flatten nested structs.

Terminology

The following list defines terms that are related to AWS Glue Data Quality.

Data Quality Definition Language (DQDL)

> A domain-specific language that you can use to write AWS Glue Data Quality rules.
>
> To learn more about DQDL, see the Data Quality Definition Language (DQDL) reference guide.

data quality

> Describes how well a dataset serves its specific purpose. AWS Glue Data Quality evaluates rules against a dataset to measure data quality. Each rule checks for particular characteristics like data freshness or integrity. To quantify data quality, you can use a *data quality score*.

data quality score

> The percentage of data quality rules that pass (result in true) when you evaluate a ruleset with AWS Glue Data Quality.

rule

A DQDL expression that checks your data for a specific characteristic and returns a Boolean value. For more information, see Rule structure.

analyzer

A DQDL expression that gathers data statistics. An analyzer gathers data statistics that can be used by ML algorithms to detect anomalies and hard-to-detect data quality issues over time.

ruleset

An AWS Glue resource that comprises a set of data quality rules. A ruleset must be associated with a table in the AWS Glue Data Catalog. When you save a ruleset, AWS Glue assigns an Amazon Resource Name (ARN) to the ruleset.

data quality score

The percentage of data quality rules that pass (result in true) when you evaluate a ruleset with AWS Glue Data Quality.

observation

An unconfirmed insight generated by AWS Glue by analyzing data statistics gathered from rules and analyzers over time.

Limits

AWS Glue Data Quality service limits:

- You can have 2,000 rules in a ruleset. If your rulesets are larger, we recommend splitting into multiple rulesets.
- The size of the ruleset is 65KB. If your rulesets are larger, we recommend splitting into multiple rulesets.
- AWS Glue Data Quality collects statistics when you create a rule or analyzer. There is no cost associated with storing these statistics. However, there is a limit of 100,000 statistics per account, and these statistics will be retained for a maximum of two years.

Release notes for AWS Glue Data Quality

This topic describes features introduced in AWS Glue Data Quality.

General availability: new features

The following new features are available with the general availability of AWS Glue Data Quality:

- The ability to identify which records failed data quality checks is now supported in AWS Glue Studio
- New data quality ruletypes such as validating referential integrity of data between two data sets, comparing data between two datasets, and data type checks
- Improved user experience in the AWS Glue Data Catalog
- Support for Apache Iceberg, Apache Hudi and Delta Lake
- Support for Amazon Redshift
- Simplified notification with Amazon EventBridge
- AWS CloudFormation support for creating rulesets
- Performance improvements: caching option in ETL and AWS Glue Studio for faster performance when evaluating data quality

Nov 27, 2023 (Preview)

- ML-powered anomaly detection capabilities are now available in AWS Glue ETL and AWS Glue Studio. With this, you can now detect anomalies and hard-to-detect data quality issues
- Dynamic Rules allows you to provide dynamic thresholds (ex: `RowCount> avg(last(10))`)

Mar 12, 2024

- DQDL improvements
  - Support for Keywords like NULL, BLANKS, WHITESPACES_ONLY
  - Options to specify how AWS Glue Data Quality must handle Composite rules
  - ColumnValues rule type will not allow NULL values to pass during comparisons
  - Support for NOT operator in DQDL

June 26, 2024

- DQDL improvements

  - DQDL now supports where clause so that you can filter data before applying DQ rules

August 7, 2024

- Anomaly Detection and Dynamic Rules are now generally available

Nov 22, 2024

- Complex composite rules allows you to author more complex business rules with nested support
- New rule types for managing data quality for your files
  - FileFreshness
  - FileSize
  - FileUniqueness
  - FileMatch
- Default data quality checks in Visual ETL jobs

Dec 6, 2024

- AWS Glue Data Quality now supports Amazon SageMaker AI LakeHouse tables and AWS Lake Formation managed Iceberg, Delta and HUDI tables in AWS Glue ETL 5.0.

Jul 7, 2025

- AWS Glue Data Quality; now supports Amazon S3 Tables, RMS, Lakehouse and AWS Lake Formation managed Iceberg tables in AWS Glue Data Catalog.

Anomaly detection in AWS Glue Data Quality

Case for Machine Learning in Data Quality

Engineers manage hundreds of data pipelines simultaneously. Each pipeline can extract data from various sources and load it into the data lake or other data repositories. To ensure high-quality data is delivered for decision-making, they establish data quality rules. These rules assess the data based on fixed criteria reflecting the current business state. However, when the business environment changes, data properties shift, rendering these fixed criteria outdated and causing poor data quality.

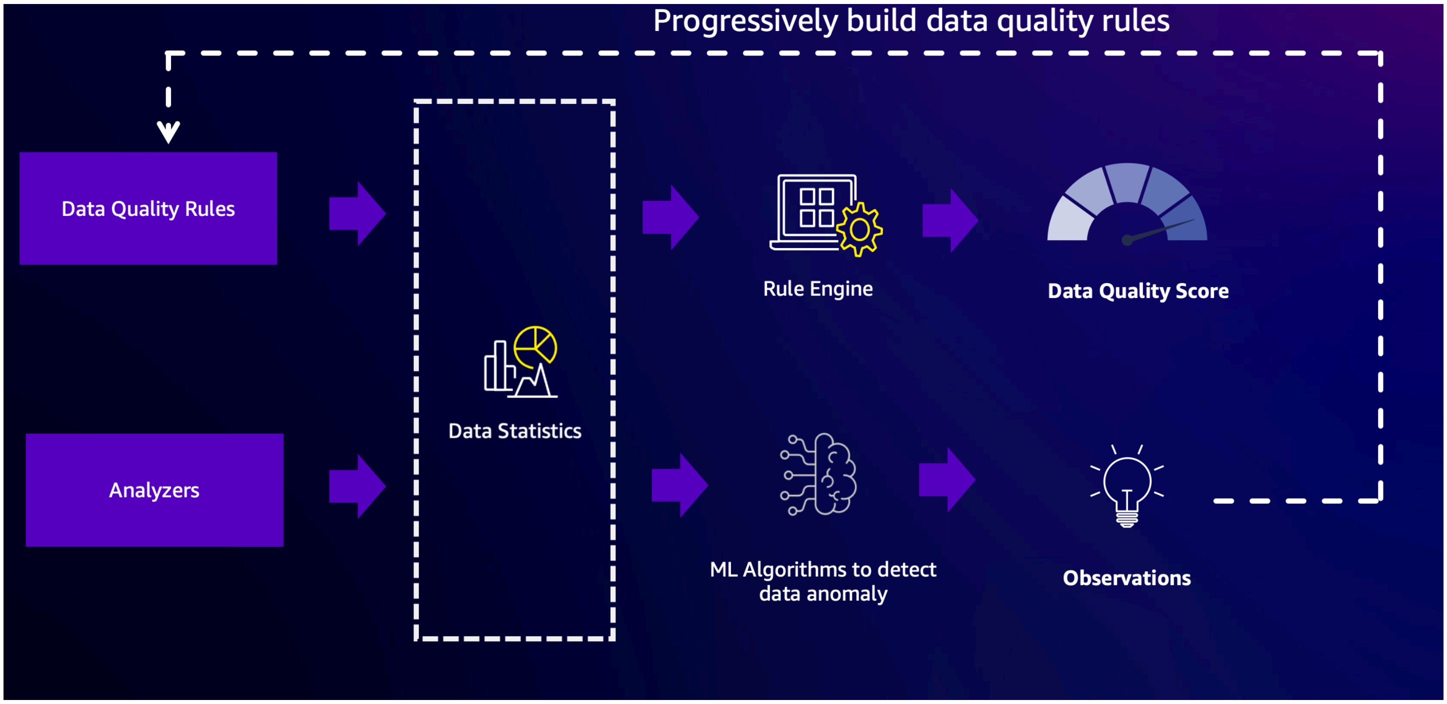
For example, a data engineer at a retail company established a rule that validates daily sales must exceed a one-million-dollar threshold. After a few months, daily sales surpassed two million dollars, rendering the threshold obsolete. The data engineer could not update the rules to reflect the latest thresholds due to a lack of notification and the effort required to manually analyze and update the rule. Later in the month, business users noticed a 25% drop in their sales. After hours of investigation, the data engineers discovered that an ETL pipeline responsible for extracting data from some stores had failed without generating errors. The rule with outdated thresholds continued to operate successfully without detecting this issue.

Alternatively, proactive alerts that can detect these anomalies could have enabled users to detect this issue. Moreover, tracking seasonality in business can highlight significant data quality problems. For instance, retail sales may be highest on weekends and during the holiday season while relatively low on weekdays. Divergence from this pattern may indicate data quality issues or shifts in business circumstances. Data quality rules cannot detect seasonal patterns as this requires advanced algorithms that can learn from past patterns capturing seasonality to detect deviations.

Finally, users find it challenging to create and maintain rules due to the technical nature of the rule creation process and the time involved to author them. As a result, they prefer exploring data insights first before defining rules. Customers need the ability to spot anomalies with ease, enabling them to proactively detect data quality issues and make confident business decisions.

How it works

> **Note**
>
> Anomaly detection is only supported in AWS Glue ETL. This is not supported in Data Catalog-based data quality.


Progressively build data quality rules
Data Quality Rules
Analyzers
Data Statistics
Rule Engine
Data Quality Score
ML Algorithms to detect data anomaly
Observations

AWS Glue Data Quality combines the power of rule-based data quality and anomaly detection capabilities to deliver high-quality data. To get started, you must first configure rules and analyzers, and then enable anomaly detection.

Rules

Rules – Rules express the expectations for your data in an open language called the Data Quality Definition Language (DQDL). An example of a rule is shown below. This rule will be successful when there are no empty or NULL values in the `passenger_count` column:

```
Rules = [
    IsComplete "passenger_count"
]
```

Analyzers

In situations where you know the critical columns but may not know enough about the data to write specific rules, you can monitor those columns using Analyzers. Analyzers are a way to gather data statistics without defining explicit rules. An example of configuring Analyzers is shown below:

```
Analyzers = [
AllStatistics "fare_amount",
DistinctValuesCount "pulocationid",
RowCount
]
```

In this example, three Analyzers are configured:

1. The first Analyzer, `AllStatistics "fare_amount"`, will capture all available statistics for the `fare_amount` field.
2. The second Analyzer, `DistinctValuesCount "pulocationid"`, will capture the count of distinct values in the `pulocationid` column.
3. The third Analyzer, `RowCount`, will capture the total number of records in the dataset.

Analyzers serve as a simple way to gather relevant data statistics without specifying complex rules. By monitoring these statistics, you can gain insights into the data quality and identify potential issues or anomalies that may require further investigation or the creation of specific rules.

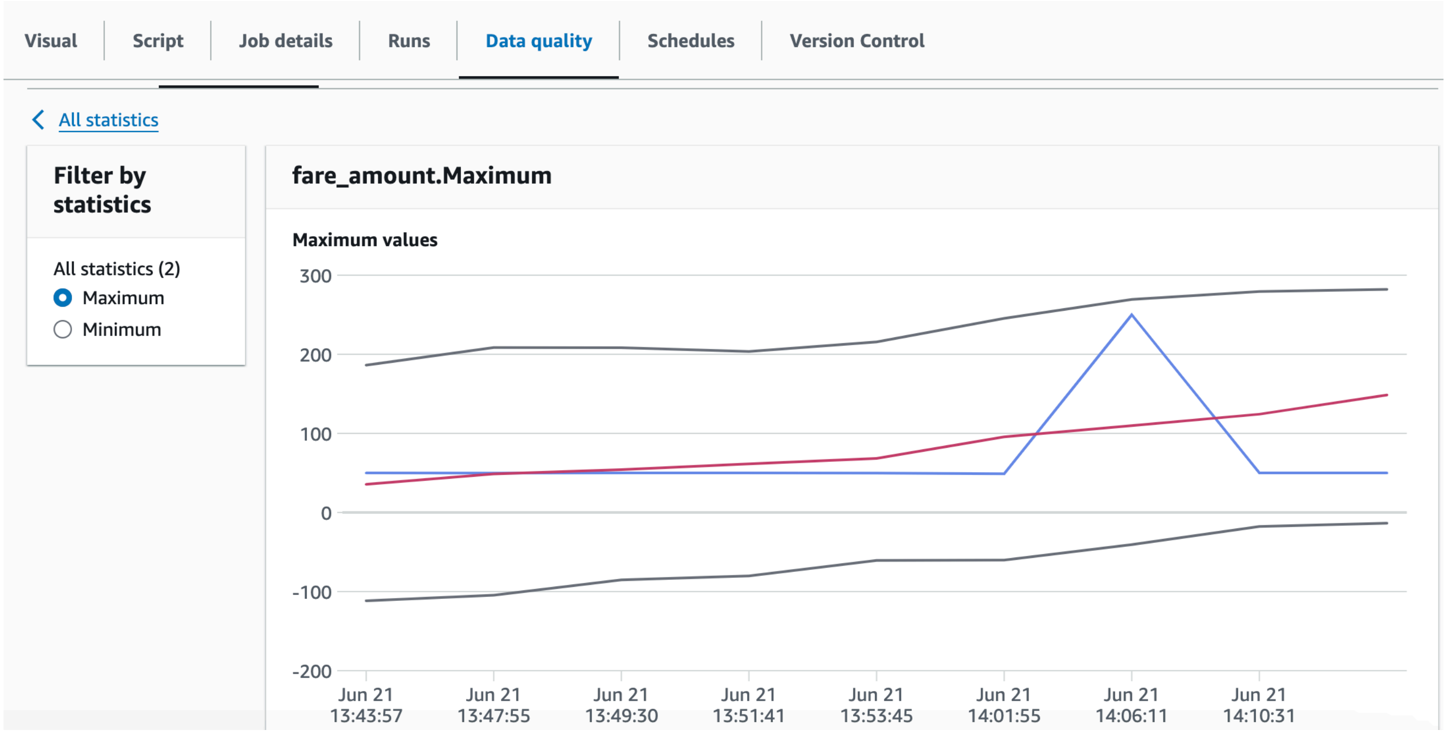
Data statistics

Both Analyzers and Rules in AWS Glue Data Quality gather data statistics, also known as data profiles. These statistics provide insights into the characteristics and quality of your data. The collected statistics are stored over time within the AWS Glue service, allowing you to track and analyze changes in your data profiles.

You can easily retrieve these statistics and write them to Amazon S3 for further analysis or long-term storage by invoking the appropriate APIs. This functionality enables you to integrate data profiling into your data processing workflows and leverage the collected statistics for various purposes, such as data quality monitoring, anomaly detection.

By storing the data profiles in Amazon S3, you can take advantage of the scalability, durability, and cost-effectiveness of Amazon's object storage service. Additionally, you can leverage other AWS services or third-party tools to analyze and visualize the data profiles, enabling you to gain deeper insights into your data quality and make informed decisions about data management and governance.

Here is an example of data statistics stored over time.

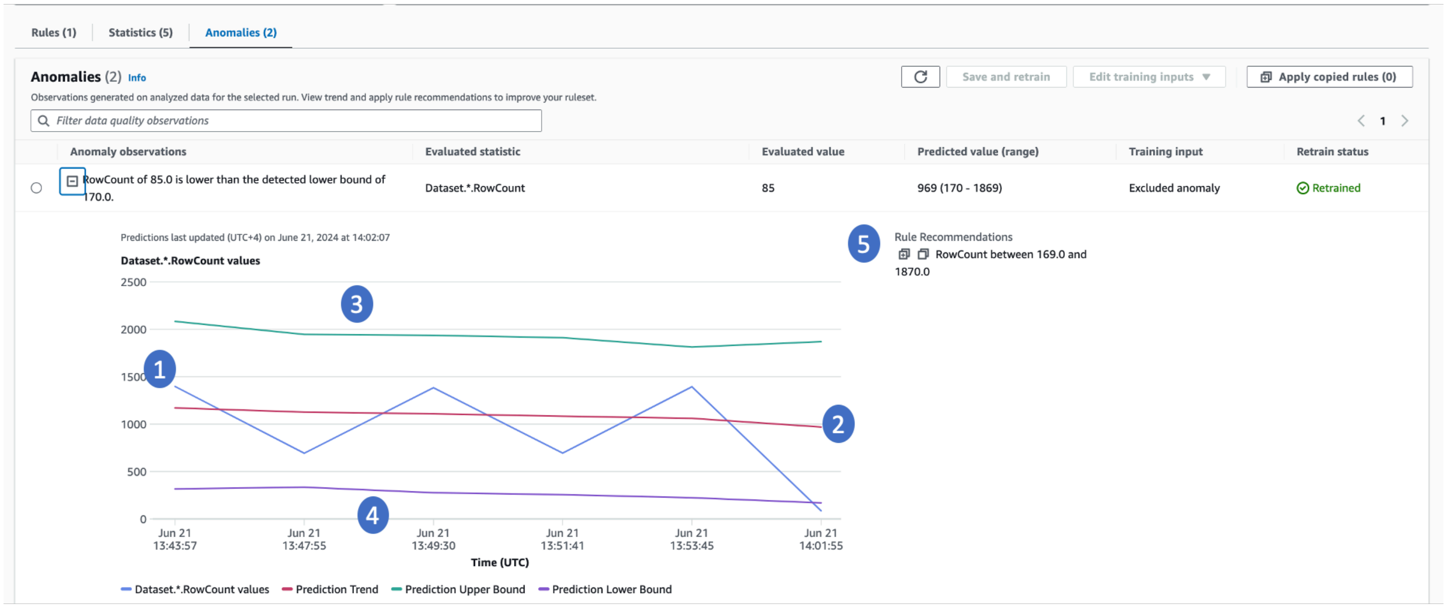

Visual
Script
Job details
Runs
Data quality
Schedules
Version Control
All statistics
Filter by statistics
All statistics (2)
● Maximum
○ Minimum
fare_amount.Maximum
Maximum values
300
200
100
0
-100
-200
Jun 21 13:43:57
Jun 21 13:47:55
Jun 21 13:49:30
Jun 21 13:51:41
Jun 21 13:53:45
Jun 21 14:01:55
Jun 21 14:06:11
Jun 21 14:10:31

Note

AWS Glue Data Quality will gather statistics only once, even if you have both **Rule** and **Analyzer** for the same columns, making the statistics generation process efficient.

Anomaly Detection

AWS Glue Data Quality requires a minimum of three data points to detect anomalies. It utilizes a machine learning algorithm to learn from past trends and then predict future values. When the actual value does not fall within the predicted range, AWS Glue Data Quality creates an Anomaly Observation. It provides a visual representation of the actual value and the trends. Four values are displayed on the graph below.


Rules (1)
Statistics (5)
Anomalies (2)
Info
Save and retrain
Edit training inputs
Apply copied rules (0)
Filter data quality observations
1
2500
2000
1500
1000
500
0
Jun 21 13:43:57
Jun 21 13:47:55
Jun 21 13:49:30
Jun 21 13:51:41
Jun 21 13:53:45
Jun 21 14:01:55
Time (UTC)
Dataset.*.RowCount values
Prediction Trend
Prediction Upper Bound
Prediction Lower Bound
1
2
3
4
5

Anomalies (2)

Observations generated on analyzed data for the selected run. View trend and apply rule recommendations to improve your ruleset.

	Anomaly observations	Evaluated statistic	Evaluated value	Predicted value (range)	Training input	Retrain status
○	RowCount of 85.0 is lower than the detected lower bound of 170.0.	Dataset.*.RowCount	85	969 (170 - 1869)	Excluded anomaly	✓ Retrained

Predictions last updated (UTC+4) on June 21, 2024 at 14:02:07

Dataset.*.RowCount values

Rule Recommendations

RowCount between 169.0 and 1870.0

1. The actual statistic and its trend over time.
2. A derived trend by learning from the actual trend. This is useful for understanding the trend direction.
3. The possible upper bound for the statistic.
4. The possible lower bound for the statistic.
5. Recommended data quality rules that can detect these issues in the future.

There are a few important things to note regarding Anomalies:

- When anomalies are generated, data quality scores are not impacted.
- When an anomaly is detected, it is considered normal for subsequent runs. The machine learning algorithm will consider this anomalous value as input unless it is explicitly excluded.

Retraining

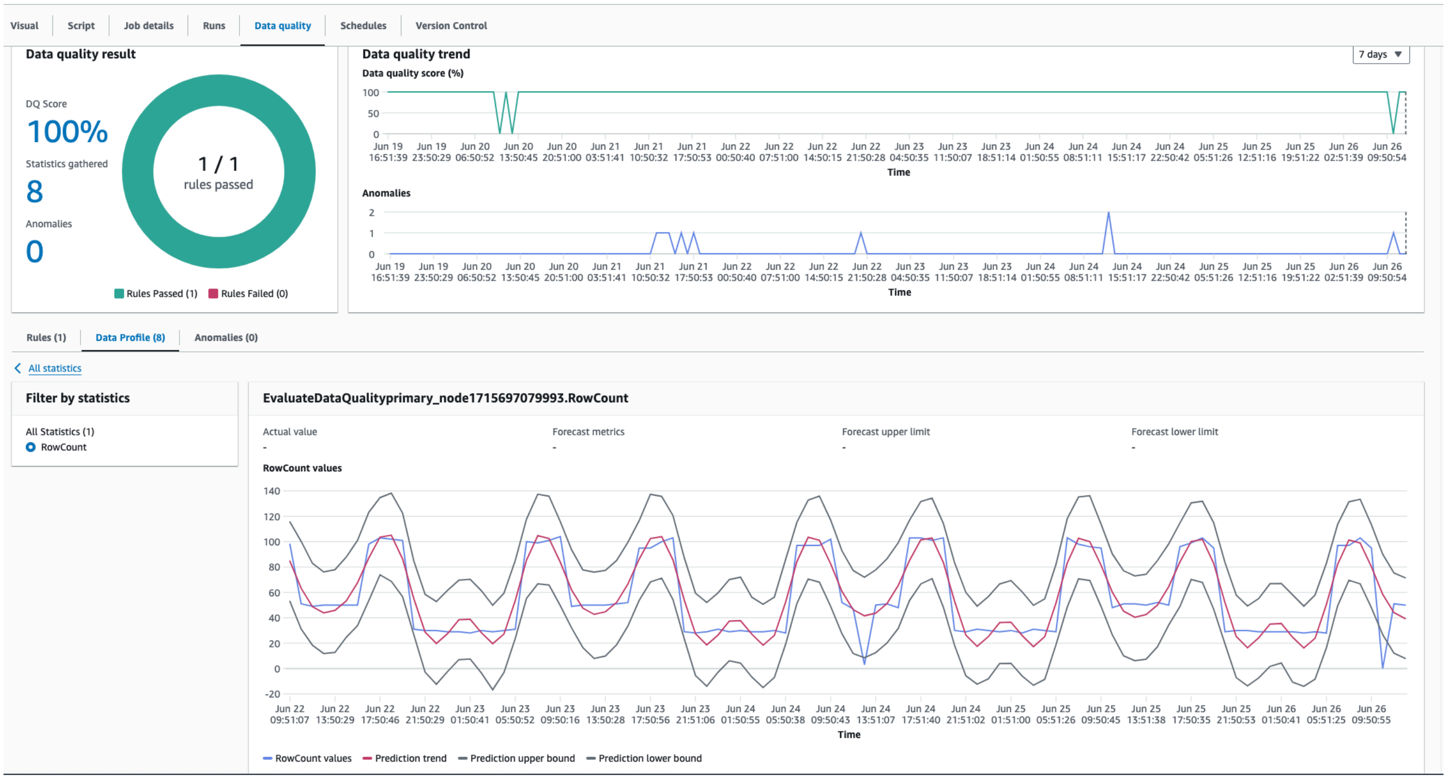
Retraining the anomaly detection model is critical to detect the right anomalies. When anomalies are detected, AWS Glue Data Quality includes the anomaly in the model as a normal value. To ensure the anomaly detection works accurately, it is important to provide feedback by acknowledging or rejecting the anomaly. AWS Glue Data Quality provides mechanisms both in AWS Glue Studio and APIs to provide feedback to the model. To know more, refer to the documentation on setting up Anomaly Detection in AWS Glue ETL pipelines.

Details of the Anomaly Detection algorithm

- The Anomaly Detection algorithm examines data statistics over time. The algorithm considers all available data points and ignores any statistics that are explicitly excluded.
- These data statistics are stored in the AWS Glue service, and you can provide AWS KMS keys to encrypt them. Refer to the Security Guide on how to provide AWS KMS keys to encrypt the AWS Glue Data Quality statistics.
- The time component is crucial for the Anomaly Detection algorithm. Based on past values, AWS Glue Data Quality determines the upper and lower bounds. During this determination, it considers the time component. The limits will differ for the same values over a one-minute interval, an hourly interval, or a daily interval.

Capturing seasonality

AWS Glue Data Quality's anomaly detection algorithm can capture seasonal patterns. For example, it can understand that weekday patterns differ from weekend patterns. This can be seen in the example below, where AWS Glue Data Quality detects a seasonal trend in the data values. You don't need to do anything specific to enable this capability. Over time, AWS Glue Data Quality learns seasonal trends and detects anomalies when these patterns break.


Visual
Script
Job details
Runs
Data quality
Schedules
Version Control
Data quality result

DQ Score
100%

Statistics gathered
8

Anomalies
0

1 / 1
rules passed

🟩 Rules Passed (1)
🟥 Rules Failed (0)
Data quality trend
7 days
Data quality score (%)
100
50
0
Jun 19 16:51:39
Jun 19 23:50:29
Jun 20 06:50:52
Jun 20 13:50:45
Jun 20 20:51:00
Jun 21 03:51:41
Jun 21 10:50:32
Jun 21 17:50:53
Jun 22 00:50:40
Jun 22 07:51:00
Jun 22 14:50:15
Jun 22 21:50:28
Jun 23 04:50:35
Jun 23 11:50:07
Jun 23 18:51:14
Jun 24 01:50:55
Jun 24 08:51:11
Jun 24 15:51:17
Jun 24 22:50:42
Jun 25 05:51:26
Jun 25 12:51:16
Jun 25 19:51:22
Jun 26 02:51:39
Jun 26 09:50:54
Time
Anomalies
2
1
0
Time
Rules (1)
Data Profile (8)
Anomalies (0)
All statistics
Filter by statistics

All Statistics (1)
● RowCount

EvaluateDataQualityprimary_node1715697079993.RowCount

Actual value
-

Forecast metrics
-

Forecast upper limit
-

Forecast lower limit
-

RowCount values
140
120
100
80
60
40
20
0
-20
Jun 22 09:51:07
Jun 22 13:50:29
Jun 22 17:50:46
Jun 22 21:50:29
Jun 23 01:50:41
Jun 23 05:50:52
Jun 23 09:50:16
Jun 23 13:50:28
Jun 23 17:50:56
Jun 23 21:51:06
Jun 24 01:50:55
Jun 24 05:50:38
Jun 24 09:50:43
Jun 24 13:51:07
Jun 24 17:51:40
Jun 24 21:51:02
Jun 25 01:51:00
Jun 25 05:51:26
Jun 25 09:50:45
Jun 25 13:51:38
Jun 25 17:50:35
Jun 25 21:50:53
Jun 26 01:50:41
Jun 26 05:51:25
Jun 26 09:50:55
Time
RowCount values
Prediction trend
Prediction upper bound
Prediction lower bound

Cost

You will be charged based on the time it takes to detect anomalies. Every statistic is charged 1 DPU for the time it takes to detect anomalies. Refer to AWS Glue Pricing for detailed examples.

Key considerations

There is no cost to store the statistics. However, there is a limit of 100,000 statistics per account. These statistics will be stored for maximum of two years.

Configure IAM permissions for AWS Glue Data Quality

This topic provides information to help you understand the actions and resources that you an IAM administrator can use in an AWS Identity and Access Management (IAM) policy for AWS Glue Data Quality. It also includes sample IAM policies with the minimum permissions you need to use AWS Glue Data Quality with the AWS Glue Data Catalog.

For additional information about security in AWS Glue, see Security in AWS Glue.

IAM permissions for AWS Glue Data Quality

The following table lists the permissions that a user needs in order to perform specific AWS Glue Data Quality operations. To set fine-grained authorization for AWS Glue Data Quality, you can specify these actions in the `Action` element of an IAM policy statement.

AWS Glue Data Quality actions

Action	Description	Resource types
`glue:CreateDataQualityRuleset`	Grants permission to create a data quality ruleset.	`::dataQualityRuleset/<name>`
`glue:DeleteDataQualityRuleset`	Grants permission to delete a data quality ruleset.	`::dataQualityRuleset/<name>`
`glue:GetDataQualityRuleset`	Grants permission to retrieve a data quality ruleset.	`::dataQualityRuleset/<name>`
`glue:ListDataQualityRulesets`	Grants permission to retrieve all data quality rulesets.	`::dataQualityRuleset/*`

Action	Description	Resource types
`glue:UpdateDataQualityRuleset`	Grants permission to update a data quality ruleset.	`::dataQualityRuleset/<name>`
`glue:GetDataQualityResult`	Grants permission to retrieve a data quality task run result. This IAM action also provides permissions to the following APIS: • `BatchGetDataQualityQualityResult` • `ListDataQualityStatistics` • `ListDataQualityStatisticAnnotations`	`::dataQualityRuleset/<name>`
`glue:ListDataQualityResults`	Grants permission to retrieve all data quality task run results.	`::dataQualityRuleset/*`
`glue:CancelDataQualityRuleRecommendationRun`	Grants permission to stop an in-progress data quality recommendation task run.	`::dataQualityRuleset/*`
`glue:GetDataQualityRuleRecommendationRun`	Grants permission to retrieve a data quality recommendation task run.	`::dataQualityRuleset/*`
`glue:ListDataQualityRuleRecommendationRuns`	Grants permission to retrieve all data quality recommendation task runs.	`::dataQualityRuleset/*`
`glue:StartDataQualityRuleRecommendationRun`	Grants permission to start a data quality recommendation task run.	`::dataQualityRuleset/*`

Action	Description	Resource types
`glue:CancelDataQualityRulesetEvaluationRun`	Grants permission to stop an in-progress data quality task run.	`::dataQualityRuleset/*`
`glue:GetDataQualityRulesetEvaluationRun`	Grants permission to retrieve a data quality task run.	`::dataQualityRuleset/*`
`glue:ListDataQualityRulesetEvaluationRuns`	Grants permission to retrieve all data quality task runs.	`::dataQualityRuleset/*`
`glue:StartDataQualityRulesetEvaluationRun`	Grants permission to start a data quality task run.	`::dataQualityRuleset/<name>`
`glue:PublishDataQuality`	Grants permission to publish data quality results.	`::dataQualityRuleset/<name>`
`glue:GetDataQualityModel`	Grants permission to retrieve Data Quality Model.	`::dataQualityRuleset/<name>, ::job/<name>`
`glue:GetDataQualityModelResult`	Grants permission to retrieve Data Quality Model Results.	`::dataQualityRuleset/<name>, ::job/<name>`
`glue:PutDataQualityStatisticAnnotation`	Grants permission to add annotations to Statistics. This IAM action also provides permissions to the following APIS: • `BatchPutDataQualityStatisticAnnotation`	`::dataQualityRuleset/<name>, ::job/<name>`

Action	Description	Resource types
`glue:PutDataQualityProfileAnnotation`	Grants permission to put annotations to all Statistics in a Profile.	`::dataQualityRuleset/<name>, ::job/<name>`

IAM setup required for scheduling evaluation runs

IAM permissions

To run scheduled Data Quality evaluation runs, you must add the `IAM:PassRole` action to the permissions policy.

AWS EventBridge Scheduler required permissions

Action	Description	Resource types
`iam:PassRole`	Grants permission for IAM to allow the user to pass the approved roles.	ARN of the role used to call `StartDataQualityRulesetEvaluationRun`

Without these permissions the following error occurs:

```
"errorCode": "AccessDenied"
"errorMessage": "User: arn:aws:sts::account_id:assumed-role/AWSGlueServiceRole is not
authorized to perform: iam:PassRole on resource: arn:aws:iam::account_id:role/service-
role/AWSGlueServiceRole
because no identity-based policy allows the iam:PassRole action"
```

IAM trusted entities

The AWS Glue and AWS EventBridge Scheduler services need to be listed in the trusted entities in order to create and run a scheduled `StartDataQualityEvaluationRun`.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {
        "Service": "glue.amazonaws.com"
      },
      "Action": "sts:AssumeRole"
    },
    {
      "Effect": "Allow",
      "Principal": {
        "Service": "scheduler.amazonaws.com"
      },
      "Action": "sts:AssumeRole"
    }
  ]
}
```

Example IAM policies

An IAM role for AWS Glue Data Quality needs the following types of permissions:

- Permissions for AWS Glue Data Quality operations so that you can get recommended data quality rules and run a data quality task against a table in the AWS Glue Data Catalog. The example IAM policies in this section include the minimum permissions required for AWS Glue Data Quality operations.
- Permissions that grant access to your Data Catalog table and the underlying data. These permissions vary depending on your use case. For example, for data that you catalog in Amazon S3, the permissions should include access to Amazon S3.

> **Note**
>
> You must configure Amazon S3 permissions in addition to the permissions described in this section.

Minimum permissions to get recommended data quality rules

This example policy includes the permissions you need in order to generate recommended data quality rules.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "AllowGlueRuleRecommendationRunActions",
      "Effect": "Allow",
      "Action": [
        "glue:GetDataQualityRuleRecommendationRun",
        "glue:PublishDataQuality",
        "glue:CreateDataQualityRuleset"
      ],
      "Resource": "arn:aws:glue:us-east-1:111122223333:dataQualityRuleset/*"
    },
 {
     "Sid": "AllowCatalogPermissions",
     "Effect": "Allow",
     "Action": [
        "glue:GetPartitions",
        "glue:GetTable"
      ],
      "Resource": [
        "*"
      ]
    },
    {
      "Sid": "AllowS3GetObjectToRunRuleRecommendationTask",
      "Effect": "Allow",
      "Action": [
        "s3:GetObject"
      ],
      "Resource": "arn:aws:s3:::aws-glue-*"
    },
  {
   "Sid": "AllowPublishingCloudwatchLogs",
   "Effect": "Allow",
   "Action": [
      "logs:CreateLogStream",
      "logs:CreateLogGroup",
      "logs:PutLogEvents"
    ],
    "Resource": "*"
     }
   ]
}
```

Minimum permissions to run a data quality task

This example policy includes the permissions you need in order to run a data quality evaluation task.

The following policy statements are optional, depending on your use case:

- `AllowCloudWatchPutMetricDataToPublishTaskMetrics` - Required if you want to publish data quality run metrics to Amazon CloudWatch.
- `AllowS3PutObjectToWriteTaskResults` - Required if you want to write data quality run results to Amazon S3.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "AllowGlueGetDataQualityRuleset",
      "Effect": "Allow",
      "Action": [
        "glue:GetDataQualityRuleset"
      ],
      "Resource": "arn:aws:glue:us-east-1:111122223333:dataQualityRuleset/YOUR-RULESET-NAME"
    },
    {
      "Sid": "AllowGlueRulesetEvaluationRunActions",
      "Effect": "Allow",
      "Action": [
        "glue:GetDataQualityRulesetEvaluationRun",
        "glue:PublishDataQuality"
      ],
      "Resource": "arn:aws:glue:us-east-1:111122223333:dataQualityRuleset/*"
    },
    {
      "Sid": "AllowCatalogPermissions",
      "Effect": "Allow",
      "Action": [
        "glue:GetPartitions",
        "glue:GetTable"
      ],
      "Resource": [
        "*"
      ]
    },
    {
      "Sid": "AllowS3GetObjectForRulesetEvaluationRun",
      "Effect": "Allow",
      "Action": [
        "s3:GetObject"
      ],
      "Resource": "arn:aws:s3:::aws-glue-*"
    },
    {
      "Sid": "AllowCloudWatchPutMetricDataToPublishTaskMetrics",
      "Effect": "Allow",
      "Action": [
        "cloudwatch:PutMetricData"
      ],
      "Resource": "*",
      "Condition": {
        "StringEquals": {
          "cloudwatch:namespace": "Glue Data Quality"
        }
      }
    },
    {
      "Sid": "AllowS3PutObjectToWriteTaskResults",
      "Effect": "Allow",
      "Action": [
        "s3:PutObject*"
      ],
      "Resource": "arn:aws:s3:::amzn-s3-demo-bucket/*"
    }
```

```
  ]
}
```

Minimum permissions to run an data quality ETL job

This example policy includes the permissions you need in order to run an data quality ETL Job.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "AllowGluePublishDataQualityResult",
      "Effect": "Allow",
      "Action": [
        "glue:PublishDataQuality"
      ],
      "Resource": "arn:aws:glue:us-east-1:111122223333:dataQualityRuleset/*"
    },
    {
      "Sid": "AllowGlueGetDataQualityResult",
      "Effect": "Allow",
      "Action": [
        "glue:GetDataQualityResult"
      ],
      "Resource": "arn:aws:glue:us-east-1:111122223333:dataQualityRuleset/*"
    },
    {
      "Sid": "AllowGlueDataQualityStatisticAnnotation",
      "Effect": "Allow",
      "Action": [
        "glue:PutDataQualityStatisticAnnotation"
      ],
      "Resource": [
        "arn:aws:glue:us-east-1:111122223333:dataQualityRuleset/*",
        "arn:aws:glue:us-east-1:111122223333::job/{JobName}"
      ]
    },
    {
      "Sid": "AllowGlueDataQualityProfileAnnotation",
      "Effect": "Allow",
      "Action": [
        "glue:PutDataQualityProfileAnnotation"
      ],
      "Resource": [
        "arn:aws:glue:us-east-1:111122223333:dataQualityRuleset/*",
        "arn:aws:glue:us-east-1:111122223333::job/{JobName}"
      ]
    }
 ]
}
```

Getting started with AWS Glue Data Quality for the Data Catalog

This getting started section provides instructions to help you get started with AWS Glue Data Quality on the AWS Glue console. You'll learn how to complete essential tasks such as generating data quality rule recommendations and evaluating a ruleset against your data.

Topics

- Prerequisites
- Step-by-step example
- Generating rule recommendations
- Monitoring rule recommendations
- Editing recommended rulesets
- Creating a new ruleset
- Running a ruleset to evaluate data quality
- Viewing the data quality score and results
- Supported source types
- Related topics

Prerequisites

Before you use AWS Glue Data Quality, you should be familiar with using the Data Catalog and crawlers in AWS Glue. With AWS Glue Data Quality, you can evaluate quality for tables in a Data Catalog database. You also need the following:

- A table in the Data Catalog to evaluate your data quality ruleset against.
- An IAM role for AWS Glue that you supply when you generate rule recommendations or run a data quality task. This role must have permission to access resources that various AWS Glue Data Quality processes require to run on your behalf. These resources include AWS Glue, Amazon S3, and CloudWatch. To view example policies that include the minimum permissions for AWS Glue Data Quality, see Example IAM policies.

  To learn more about IAM roles for AWS Glue, see Create an IAM policy for the AWS Glue service and Create an IAM role for the AWS Glue service. You can also view a list of all AWS Glue permissions that are specific to data quality in Authorization for AWS Glue Data Quality actions.
- A database with at least one table that contains a variety of data. The table used in this tutorial is named `yyz-tickets`, with the table `tickets`. This data is a collection of publicly available information from the City of Toronto for parking citations. If you create your own table, make sure that it's populated with a variety of valid data to get the best set of recommended rules.

Step-by-step example

For a step-by-step example with sample datasets, see the AWS Glue Data Quality blog post.

Generating rule recommendations

Rule recommendations make it easy to get started with data quality without writing code. With AWS Glue Data Quality, you can analyze your data, identify rules, and create a ruleset that you can evaluate in a data quality task. Recommendation runs are automatically deleted after 90 days.

To generate data quality rule recommendations

1. Open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. Choose **Tables** in the navigation pane. Then choose the table that you want to generate data quality rule recommendations for.
3. On the table details page, select the **Data quality** tab to access AWS Glue Data Quality rules and settings for your table.
4. On the **Data quality** tab, choose **Add rules and monitor data quality**.
5. On the **Ruleset builder** page, an alert at the top of the page will prompt you to start a recommendation task if there are no rule recommendation runs.

6. Choose **Recommend rules** to open the modal and input your parameters for the recommendation task.
7. Choose an IAM role with access to AWS Glue. This role must have permission to access resources that various AWS Glue Data Quality processes require to run on your behalf.
8. After the fields are completed according to your preferences, choose **Recommend rules** to start the recommendation task run. If recommendation runs are in progress or completed, you can manage your runs in this alert. You might need to refresh the alert to view the status change. Completed and in-progress recommendation task runs appear in the **Run history** page that lists all recommendation runs for the past 90 days.

What the recommended rules mean

AWS Glue Data Quality generates rules based on the data from each column of the input table. It uses the rules to identify potential boundaries where data can be filtered to maintain quality requirements. The following list of generated rules includes examples that are useful for understanding what the rules mean and what they might do when applied to your data.

For a full list of the generated Data Quality Definition Language (DQDL) rule types, see DQDL rule type reference.

- `IsComplete "SET_FINE_AMOUNT"` –The `IsComplete` rule verifies that the column is filled in for any given row. Use this rule to tag columns as non-optional in data.
- `Uniqueness "TICKET_NUMBER" > 0.95` – The `Uniqueness` rule verifies that the data within the column meets some threshold of uniqueness. In this example, the data that populates any given row for `"TICKET_NUMBER"` was determined to be at most 95% identical in content to all other rows, which suggests this rule.
- `ColumnValues "PROVINCE" in ["ON", "QC", "AB", "NY",...]` – The `ColumnValues` rule defines valid values for the column, based on existing column contents. In this example, the data for each row is a 2-letter license code plate for a state or province.
- `ColumnLength "INFRACTION_DESCRIPTION" between 15 and 31` – The `ColumnLength` rule enforces a length restriction on a column's data. This rule is generated from the sample data based on the minimum and maximum recorded lengths for a column of strings.

Monitoring rule recommendations

When data quality rule recommendations are running, the **Add rules and monitor data quality** page displays information and additional actions that you can take in the top bar.

When rule recommendations are in progress, you can choose **Stop run** before the recommendation task is complete. While the task is in progress, you will see the status, **in progress**, and the date and time when the run started.

When the rule recommendations are complete, the rule recommendation bar displays the number of rules recommended, the status of the last recommendation run, and the date and timestamp when it finished.

You can add the recommended rules by choosing **Insert Rule Recommendation**. To view previously recommended rules, select a specific date. To run a new recommendation, choose **More actions**, and then choose **Recommended rules**.

Set default settings by choosing **Manage user settings**. You can set the default path for Amazon S3 to store rulesets or to set up a default role to run the Data Catalog.

Editing recommended rulesets

Because AWS Glue Data Quality generates rules based on existing data that you have available, you might see some unexpected or undesirable rules in the automated suggestions. In order to get the most out of the recommended rulesets, you need to evaluate and modify them. For this step of the tutorial, you take the rules generated in the previous step and adjust them to enforce more restrictive qualities on some data. You also relax other rules to ensure that correct, unique data can be added later.

Edit a suggested ruleset

1. In the AWS Glue console, choose **Data Catalog**, and then choose **Databases tables** in the navigation pane. Choose the table `tickets`.
2. On the table details page, choose the **Data quality** tab to access AWS Glue Data Quality rules and settings for the table.
3. In the **Rulesets** section, select the ruleset generated in Generating rule recommendations.
4. Choose **Actions**, and then choose **Edit** in the console window. The ruleset editor loads in the console. It includes an editing pane for your rules and a quick reference for DQDL.

5. Remove line 2 of the script. This relaxes the requirement that the database size is constrained within a certain number of rows. After the edit, your file should contain the following on lines 1–3:

```
Rules = [
    IsComplete "TAG_NUMBER_MASKED",
    ColumnLength "TAG_NUMBER_MASKED" between 6 and 9,
```

6. Remove line 25 of the script. This relaxes the requirement that 96% of recorded provinces are `ON`. After the edit, your file should contain the following from line 24 to the end of the ruleset:

```
ColumnValues "PROVINCE" in ["ON", "QC", "AB", "NY", "AZ", "NS", "BC", "MI", "PQ",
 "MB", "PA", "FL", "SK", "NJ", "OH", "NB", "IL", "MA", "CA",
    "VA", "TX", "NF", "MD", "PE", "CT", "NC", "GA", "IN", "OR", "MN", "TN", "WI",
 "KY", "MO", "WA", "NH", "SC", "CO", "OK", "VT", "RI", "ME", "AL",
    "YT", "IA", "DE", "AR", "LA", "XX", "WV", "MT", "KS", "NT", "DC", "NV", "NE",
 "UT", "MS", "NM", "ID", "SD", "ND", "AK", "NU", "GO", "WY", "HI"],
ColumnLength "PROVINCE" = 2
]
```

7. Change line 14 to the following:

```
IsComplete "TIME_OF_INFRACTION",
```

   This *strengthens* the requirement on the column by limiting the database to only tickets that contain a recorded time of infraction. You should always consider tickets without a recorded time of infraction to be invalid data in this dataset. This is different than situations where partitioning or transformation might be more appropriate for further data use or inspection to determine a quality rule.

8. Choose **Update Ruleset**at the bottom of the console page.

Creating a new ruleset

A ruleset is a group of data quality rules that you evaluate against your data. In the AWS Glue console, you can author custom rulesets using Data Quality Definition Language (DQDL).

To create a data quality ruleset

1. In the AWS Glue console, choose **Data Catalog**, choose **Databases**, and then choose **Tables** in the navigation pane. Select the table `tickets`.
2. Open the **Data quality** tab.
3. In the **Rulesets** section, choose **Create ruleset**. The DQDL editor launches in the console. It has a text area for direct editing, and a quick reference for DQDL rules and the table schema.
4. Start adding rules to the text area of the DQDL editor. You can either write rules directly from this tutorial, or use the **DQDL rule builder** feature of the data quality rules editor.

   > **Note**
   >
   > **How to use the DQDL rule builder**
   >
   > 1. Select a rule type from the list, and select the plus sign to insert example syntax into the editor pane.
   > 2. Exchange the placeholder column names with your own column names. Column names from the table are available in the **Schema** tab.
   > 3. Update the expression parameter as you see fit. For a full list of expressions that DQDL supports, see Expressions.

   As an example, the following rules are constraints for data validation of the `ticket_number` column in the `tickets` table. To add the following rules, use the DQDL rule builder or directly edit your ruleset:

   ```
   IsComplete "ticket_number",
   IsUnique "ticket_number",
   ColumnValues "ticket_number" > 9000000000
   ```
5. Provide a name for your new ruleset in the **Ruleset name** field.
6. Choose **Save ruleset**.

Evaluating data quality across multiple datasets

You can set up data quality rules across multiple datasets using ReferentialIntegrity and DatasetMatch rulesets. ReferentialIntegrity checks to see if data in the primary dataset is present in other datasets.

To add a reference dataset, choose the **Schema** tab and then choose **Update reference tables**. You will be prompted to select a database and a table. You can add the table and then set up data quality rules. Rule types like AggregateMatch, RowCountMatch, ReferentialIntegrity, SchemaMatch, and DatasetMatch support the ability to perform data quality checks across multiple datasets.

Running a ruleset to evaluate data quality

When you run a data quality task, AWS Glue Data Quality evaluates a ruleset against your data and calculates a data quality score. This score represents the percentage of data quality rules that passed for the input.

To run a data quality task

1. In the AWS Glue console, choose **Data Catalog**, choose **Databases**, and then choose **Tables** in the navigation pane. Select the table `tickets`.
2. Choose the **Data quality** tab.
3. In the **Rulesets** list, select the ruleset that you want to evaluate against the table. For this step, we recommend using a ruleset that you've written or modified already rather than generated rules. Choose **Run**.
4. In the modal, choose your IAM role. This role must have permission to access resources that various AWS Glue Data Quality processes require to run on your behalf. You can save the IAM role as the default or modify it by going to the **Default Setting** page.
5. Under **Data quality actions**, choose whether you want to **Publish metrics to Amazon CloudWatch**. When this option is selected, AWS Glue Data Quality publishes metrics that indicate the number of rules that passed and the number of rules that failed. To take action on metrics stored this way, you can use CloudWatch alarms. Key metrics are also published to Amazon EventBridge for you to set up alerts. For more information, see Setting up alerts, deployments, and scheduling.
6. In **Run Frequency**, choose run on demand or schedule the ruleset. When you schedule a ruleset, you're prompted for a task name. The schedule will be created in Amazon EventBridge. You can edit your schedule in Amazon EventBridge.

7. To save the data quality results in Amazon S3, select a **Data quality results location**. The IAM role that you previously selected for this task must have write access to this location.
8. Under **Additional Configurations**, enter the **Requested number of workers** that you want AWS Glue to allocate for your data quality task.
9. You can optionally set up a filter at the data source. This helps you reduce the data that you're reading. You can also use a filter to run incremental validations by selecting partition information and passing them as parameters via API calls. To improve performance, you can provide a partition predicate.
10. Choose **Run**. You should see your new task in the **Data quality task runs** list. When the **Run status** column for the task shows as **Completed**, you can view the quality score results. You might need to refresh your console window for the status to update correctly.
11. To view the column for the data quality result details, choose the “+” icon to expand the ruleset. The results show you the rules that passed and failed in the evaluation, and what triggered the rule failure.

Viewing the data quality score and results

To see the latest run on all created rulesets

1. In the AWS Glue console, choose **Tables** in the navigation pane. Then choose the table that you want to run a data quality task for.
2. Choose the **Data quality** tab.
3. The **Data quality snapshot** shows a general trend of runs over time. The last 10 runs over all rulesets are displayed by default. To filter by ruleset, select the desired one from the dropdown list. If there are less than 10 runs, all the completed runs that are available are displayed.
4. In the **Data quality** table, each ruleset with its latest run (if there is one) is shown, along with the score. Expanding the ruleset displays the rules that are in that ruleset, along with the rule results for that run.

To see the latest run on a particular ruleset

1. In the AWS Glue console, choose **Tables** in the navigation pane. Then choose the table that you want to run a data quality task for.
2. Choose the **Data quality** tab.

3. In the **Data quality** table, choose on a specific ruleset.
4. On the **Ruleset details** page, choose the **Run history** tab.

   All of the evaluation runs for this particular ruleset are listed in the table within this tab. You can see the history of the scores and the status of the runs.
5. To see more information about a particular run, choose the **Run ID** to go to the **Evaluation run details** page. On this page, you can see specifics about the run and more details about the status of individual rule results.

Supported source types

Table type support by AWS Lake Formation configuration

Table Type	AWS Lake Formation - All Table Access	AWS Lake Formation Enabled with Columns	AWS Lake Formation Enabled with Data Filters	Cross-Account AWS Lake Formation Support - All Table Access	AWS Lake Formation Disabled
Parquet	Supported	Not Supported	Not Supported	Supported	Supported
ORC	Supported	Not Supported	Not Supported	Supported	Supported
CSV, JSON, TSV	Supported	Not Supported	Not Supported	Supported	Supported
Avro	Supported	Not Supported	Not Supported	Supported	Supported
JSON	Supported	Not Supported	Not Supported	Supported	Supported
Iceberg	Supported	Not Supported	Not Supported	Supported	Supported

Table Type	AWS Lake Formation - All Table Access	AWS Lake Formation Enabled with Columns	AWS Lake Formation Enabled with Data Filters	Cross-Account AWS Lake Formation Support - All Table Access	AWS Lake Formation Disabled
HUDI	Not Supported	Not Supported	Not Supported	Not Supported	Supported
Delta	Not Supported	Not Supported	Not Supported	Not Supported	Supported
RMS	Supported*	Supported*	Supported*	Not Supported	Not Supported
Amazon S3 Tables	Supported*	Not Supported	Not Supported	Not applicable	Supported
Amazon RDS and Aurora	Not applicable	Not applicable	Not applicable	Not applicable	Not Supported
JDBC	Not applicable	Not applicable	Not applicable	Not applicable	Supported

* Amazon S3 tables and SageMaker Lakehouse support in AWS Glue Console is not supported. Currently, Amazon S3 table and SageMaker Lakehouse Data Catalog Recommendation Runs and Data Catalog Data Quality Evaluation runs are only supported via the CLI.

Other known limitations

- Delta Lake Symlink Tables: Not supported for AWS Glue Data Quality recommendation runs or Data Catalog Data Quality evaluation runs.
- Amazon S3 Table Asset publishing in SageMaker Unified Studio: Currently, publishing Amazon S3 Tables as Assets in SageMaker Unified Studio is unavailable; visualizing Amazon S3 Table Data Quality runs is unavailable from SageMaker Unified Studio as a result.

Related topics

- DQDL rule type reference
- Data Quality Definition Language (DQDL) reference

Evaluating data quality with AWS Glue Studio

AWS Glue Data Quality evaluates and monitors the quality of your data based on rules that you define. This makes it easy to identify the data that needs action. In AWS Glue Studio, you can add data quality nodes to your visual job to create data quality rules on tables in your Data Catalog. You can then monitor and evaluate changes to your datasets as they evolve over time. For an overview of how to work with AWS Glue Data Quality in AWS Glue Studio, see the following video.

The following are the high-level steps for how you work with AWS Glue Data Quality:

1. **Create data quality rules –** Build a set of data quality rules using the DQDL builder by choosing built-in rulesets that you configure.
2. **Configure a data quality job –** Define actions based on the data quality results and output options.
3. **Save and run a data quality job –** Create and run a job. Saving the job will save the rulesets that you created for the job.
4. **Monitor and review the data quality results –** Review the data quality results after the job run is complete. Optionally, schedule the job for a future date.

Benefits

Data analysts, data engineers, and data scientists can use the Evaluate Data Quality node in AWS Glue Studio to analyze, configure, monitor, and improve the quality of data from the visual job editor. The benefits of using the data quality node include the following:

- **You can detect data quality issues** - You can check for issues by creating rules that check characteristics of your datasets.
- **It's easy to get started** - You can start with pre-built rules and actions.
- **Tight integration** - You can use data quality nodes in AWS Glue Studio because AWS Glue Data Quality runs on top of the AWS Glue Data Catalog.

Evaluating data quality for ETL jobs in AWS Glue Studio

In this tutorial, you get started with AWS Glue Data Quality in AWS Glue Studio. You will learn how to:

- Create rules using the Data Quality Definition Language (DQDL) rule builder.
- Specify data quality actions, data to output, and the output location of the data quality results.
- Review data quality results.

To practice with an example, review the blog post **Getting started with AWS Glue Data Quality for ETL pipelines**.

Step 1: Add the Evaluate Data Quality transform node to the visual job

In this step, you add the Evaluate Data Quality node to the visual job editor.

To add the data quality node

1. In the AWS Glue Studio console, choose **Visual with a source and target** from the **Create job** section, and then choose **Create**.
2. Choose a node that you want to apply the data quality transform to. Typically, this will be a transform node or a data source.
3. Open the resource panel on the left by choosing the "+" icon. Then search for **Evaluate Data Quality** in the search bar and choose **Evaluate Data Quality** from the search results.
4. The visual job editor showsthe **Evaluate Data Quality** transform node branching from the node you selected. On the right side of the console, the **Transform** tab is automatically opened. If you need to change the parent node, choose the **Node properties** tab, and then choose the node parent from the dropdown menu.

   When you choose a new node parent, a new connection is made between the parent node and the **Evaluate Data Quality** node. Remove any unwanted parent nodes. Only one parent node can be connected to one **Evaluate Data Quality** node.
5. The Evaluate Data Quality transform supports multiple parents so you can validate data quality rules across multiple datasets. Rules that support multiple datasets include ReferentialIntegrity, DatasetMatch, SchemaMatch, RowCountMatch, and AggregateMatch.

When you add multiple inputs to the Evaluate Data Quality transform, you need to select your “primary” input. Your primary input is the dataset that you want to validate data quality for. All other nodes or inputs are treated as references.

You can use the Evaluate Data Quality transform to identify specific records that failed data quality checks. We recommend that you choose your primary dataset because new columns that flag bad records are added to the primary dataset.

6. You can specify aliases for input data sources. Aliases provide another way to reference the input source when you're using the ReferentialIntegrity rule. Becauseonly one data source can be designated as the primary source, each additional data source that you add will require an alias.

   In the following example, the ReferentialIntegrity rule specifies the input data source by the alias name and performs a one-to-one comparison to the primary data source.

```
Rules = [
 ReferentialIntegrity “Aliasname.name” = 1
]
```

Step 2: Create a rule using DQDL

In this step, you create a rule using DQDL. For this tutorial, you create a single rule using the **Completeness** rule type. This rule type checks the percentage of complete (non-null) values in a column against a given expression. For more information about using DQDL, see DQDL.

1. On the **Transform** tab, add a **Rule type** by choosing the **Insert** button. This adds the rule type to the rule editor, where you can enter the parameters for the rule.

   Note

   When you're editing rules, ensure that the rules are within the brackets and that the rules are separated by commas. For example, a complete rule expression will look like the following:

```
Rules= [
    Completeness "year">0.8, Completeness "month">0.8
]
```

> This example specifies the parameter for completeness for the columns named 'year' and 'month'. For the rule to pass, these columns must be greater than 80% 'complete', or must have data in over 80% of instances for each respective column.

In this example, search for and insert the **Completeness** rule type. This adds the rule type to the rule editor. This rule type has the following syntax: `Completeness <COL_NAME> <EXPRESSION>`.

Most rule types require that you provide an expression as a parameter in order to create a Boolean response. For more information on supported DQDL expressions, see DQDL expressions. Next, you'll add the column name.

2. In the DQDL rule builder, choose the **Schema** tab. Use the search bar to locate the column name in the input schema. The input schema displays the column name and data type.
3. In the rule editor, click to the right of the rule type to insert the cursor where the column will be inserted. Alternately, you can enter in the name of the column in the rule.

   For example, from the list of columns in the input schema list, choose the **Insert** button next to the column (in this example, **year**). This adds the column to the rule.
4. Then, in the rule editor, add an expression to evaluate the rule. Because the **Completeness** rule type checks the percentage of complete (non-null) values in a column against a given expression, enter an expression such as `> 0.8`. This rule checks the column if it's greater than 80% complete (non-null) values.

Step 3: Configure data quality outputs

After creating data quality rules, you can select additional options to specify data quality node output.

1. In **Data quality transform output**, choose from the following options:

   - **Original data** – Choose to output original input data. When you choose this option, a new child node “rowLevelOutcomes” is added to the job. The schema matches the schema of the primary dataset that was passed as input to the transform. This option is useful if you just want to pass the data through and fail the job when quality issues occur.

Another use case is when you want to detect bad records that failed data quality checks. To detect bad records, choose the option **Add new columns to indicate data quality errors**. This action adds four new columns to the schema of the “rowLevelOutcomes” transform.

- **DataQualityRulesPass** (string array) – Provides an array of rules that passed data quality checks.
- **DataQualityRulesFail** (string array) – Provides an array of rules that failed data quality checks.
- **DataQualityRulesSkip** (string array) – Provides an array of rules that were skipped. The following rules cannot identify error records because they're applied at the dataset level.
  - AggregateMatch
  - ColumnCount
  - ColumnExists
  - ColumnNamesMatchPattern
  - CustomSql
  - RowCount
  - RowCountMatch
  - StandardDeviation
  - Mean
  - ColumnCorrelation
- **DataQualityEvaluationResult –** Provides “Passed” or “Failed” status at the row level. Note that your overall results can be FAIL, but a certain record might pass. For example, the RowCount rule might have failed, but all other rules might have been successful. In such instances, this field status is 'Passed'.

2. **Data quality results** – Choose to output configured rules and their pass or fail status. This option is useful if you want to write your results to Amazon S3 or other databases.
3. **Data quality output settings** (Optional) – Choose **Data quality output settings** to reveal the **Data quality result location** field. Then, choose **Browse** to search for an Amazon S3 location to set as the data quality output target.

Step 4. Configure data quality actions

You can use actions to publish metrics to CloudWatch or to stop jobs based on specific criteria. Actions are only available after you create a rule. When you choose this option, the same metrics are also published to Amazon EventBridge. You can use these options to create alerts for notification.

- **On ruleset failure** – You can choose what to do if a ruleset fails while the job is running. If you want the job to fail if data quality fails, choose when the job should fail by selecting one of the following options. By default, this action is not selected, and the job completes its run even if data quality rules fail.
  - **None** – If you choose **None** (default), the job does not fail and continues to run despite ruleset failures.
  - **Fail job after loading data to target** – The job fails and no data is saved. In order to save the results, choose an Amazon S3 location where the data quality results will be saved.
  - **Fail job without loading to target data** – This option fails the job immediately when a data quality error occurs. It does not load any data targets, including the results from the data quality transform.

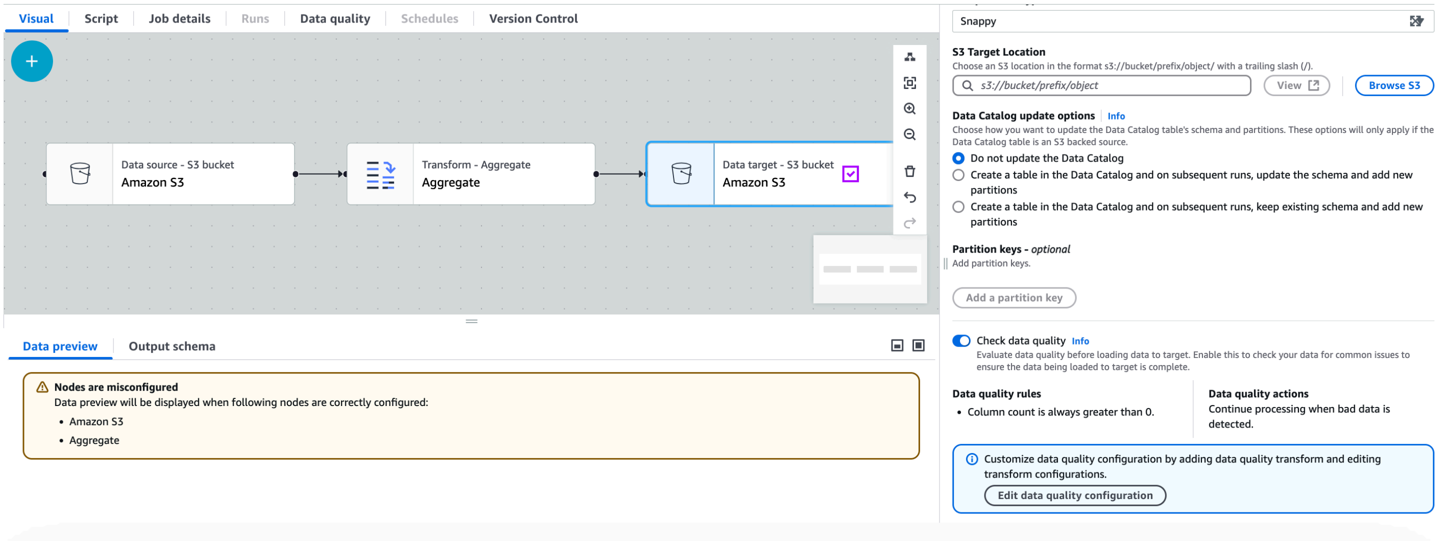
Step 5: View data quality results

After running the job, view the data quality results by choosing the **Data quality** tab.

1. For each job run, view the data quality results. Each node displays a data quality status and status detail. Choose a node to view all rules and the status of each rule.
2. Choose **Download results** to download a CSV file that contains information about the job run and data quality results.
3. If you have more than one job run with data quality results, you can filter the results by date and time range. Choose *Filter by a date and time range* to expand the filter window.
4. Choose a relative range or absolute range. For absolute ranges, use the calendar to select a date, and enter values for start time and end time. When you're done, choose **Apply**.

Automatic Data Quality

When you create a AWS Glue ETL job with Amazon S3 as the target, AWS Glue ETL automatically enables a Data Quality rule that checks if the data being loaded has at least one column. This rule is designed to ensure that the data being loaded is not empty or corrupted. However, if this rule fails, the job will not fail; instead, you will notice a reduction in your data quality score. Additionally, Anomaly Detection is enabled by default, which monitors the number of columns in the data. If there are any variations or abnormalities in the column count, AWS Glue ETL will inform you about these anomalies. This feature helps you identify potential issues with the data and take appropriate actions. To view the Data Quality rule and its configuration, you can click on the Amazon S3 target in your AWS Glue ETL job. The rule's configuration will be displayed, as shown in the provided screenshot.



Nodes are misconfigured
Data preview will be displayed when following nodes are correctly configured:

- Amazon S3
- Aggregate

Snappy

S3 Target Location
Choose an S3 location in the format s3://bucket/prefix/object/ with a trailing slash (/).
s3://bucket/prefix/object

Data Catalog update options
Choose how you want to update the Data Catalog table's schema and partitions. These options will only apply if the Data Catalog table is an S3 backed source.

● Do not update the Data Catalog
○ Create a table in the Data Catalog and on subsequent runs, update the schema and add new partitions
○ Create a table in the Data Catalog and on subsequent runs, keep existing schema and add new partitions

Partition keys - *optional*
Add partition keys.

Check data quality
Evaluate data quality before loading data to target. Enable this to check your data for common issues to ensure the data being loaded to target is complete.

Data quality rules

- Column count is always greater than 0.

Data quality actions
Continue processing when bad data is detected.

Customize data quality configuration by adding data quality transform and editing transform configurations.

You can add additional data quality rules by selecting **Edit data quality configuration**.

Aggregated Metrics

You may require aggregated metrics such as number of records that passed, failed, skipped at a rule level or at the ruleset level to build dashboards. In order to get the Aggregated Metrics and Rule Metrics for each rule, first, enable Aggregated Metrics by adding the `publishAggregatedMetrics` option to your `EvaluateDataQuality` function.

The possible options for `additional_options` `publishAggregatedMetrics` are `ENABLED` and `DISABLED`. As an example:

```
EvaluateDataQualityMultiframe = EvaluateDataQuality().process_rows(
    frame=medicare_dyf,
    ruleset=EvaluateDataQuality_ruleset,
    publishing_options={
        "dataQualityEvaluationContext": "EvaluateDataQualityMultiframe",
        "enableDataQualityCloudWatchMetrics": False,
        "enableDataQualityResultsPublishing": False,
    },
    additional_options={"publishAggregatedMetrics.status": "ENABLED"},
)
```

If not specified, the `publishAggregatedMetrics.status` is `DISABLED` by default and the ruleMetrics and aggregated metrics would now be computed. This feature is currently supported in AWS Glue Interactive Sessions and in Glue ETL jobs. This is not supported in Glue Catalog Data Quality APIs.

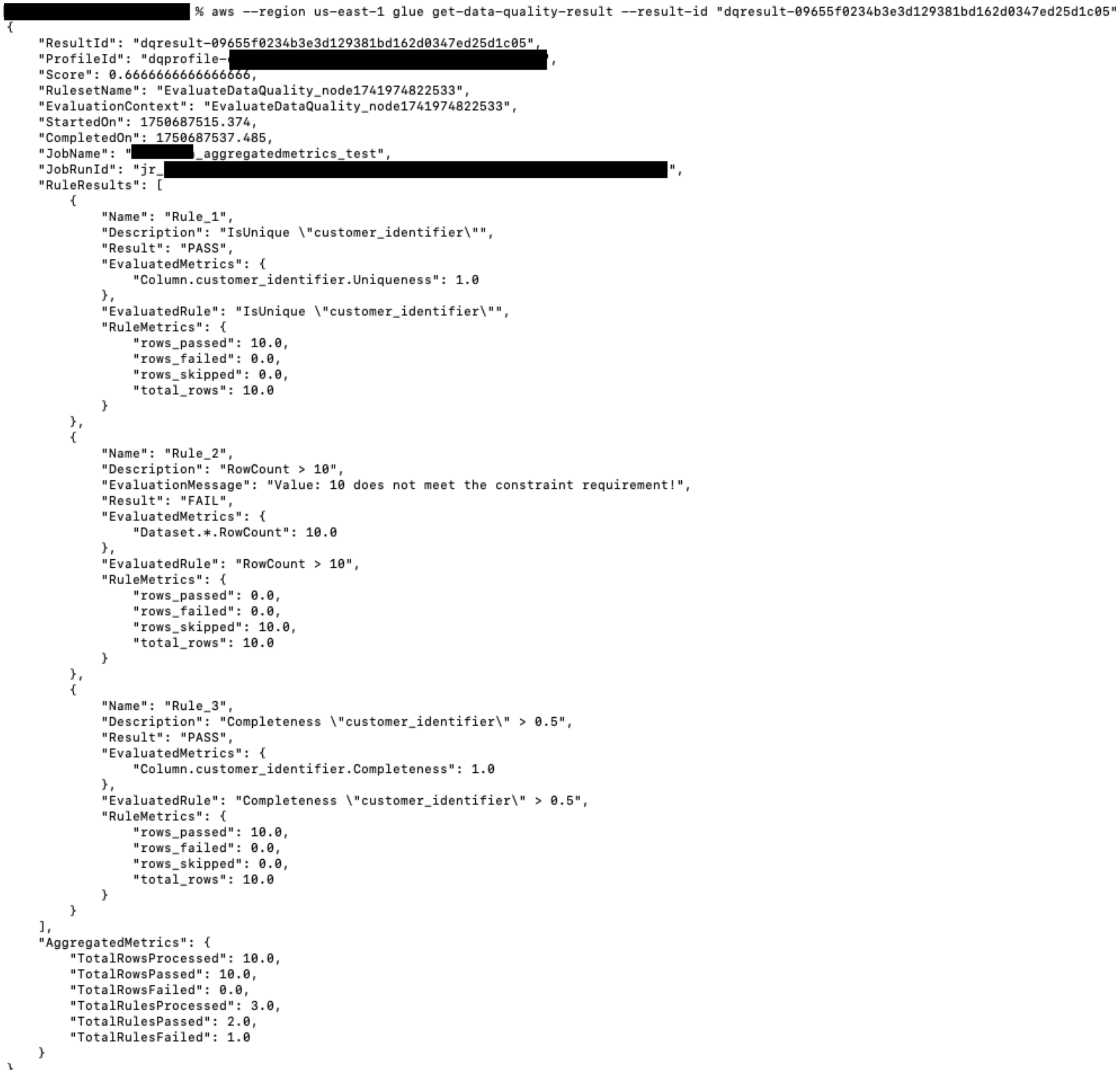
Retrieving aggregated metrics results

When `additionalOptions` `"publishAggregatedMetrics.status": "ENABLED"`, you can get the results in two places:

1. `AggregatedMetrics` and `RuleMetrics` are returned through the `GetDataQualityResult()` when providing the `resultId` where `AggregatedMetrics` and `RuleMetrics` include:

   Aggregated Metrics:

   - Total rows processed
   - Total rows passed
   - Total rows failed
   - Total Rules Processed
   - Total Rules Passed
   - Total Rules Failed



```
[redacted] % aws --region us-east-1 glue get-data-quality-result --result-id "dqresult-09655f0234b3e3d129381bd162d0347ed25d1c05"
{
    "ResultId": "dqresult-09655f0234b3e3d129381bd162d0347ed25d1c05",
    "ProfileId": "dqprofile-[redacted]",
    "Score": 0.6666666666666666,
    "RulesetName": "EvaluateDataQuality_node1741974822533",
    "EvaluationContext": "EvaluateDataQuality_node1741974822533",
    "StartedOn": 1750687515.374,
    "CompletedOn": 1750687537.485,
    "JobName": "[redacted]_aggregatedmetrics_test",
    "JobRunId": "jr_[redacted]",
    "RuleResults": [
        {
            "Name": "Rule_1",
            "Description": "IsUnique \"customer_identifier\"",
            "Result": "PASS",
            "EvaluatedMetrics": {
                "Column.customer_identifier.Uniqueness": 1.0
            },
            "EvaluatedRule": "IsUnique \"customer_identifier\"",
            "RuleMetrics": {
                "rows_passed": 10.0,
                "rows_failed": 0.0,
                "rows_skipped": 0.0,
                "total_rows": 10.0
            }
        },
        {
            "Name": "Rule_2",
            "Description": "RowCount > 10",
            "EvaluationMessage": "Value: 10 does not meet the constraint requirement!",
            "Result": "FAIL",
            "EvaluatedMetrics": {
                "Dataset.*.RowCount": 10.0
            },
            "EvaluatedRule": "RowCount > 10",
            "RuleMetrics": {
                "rows_passed": 0.0,
                "rows_failed": 0.0,
                "rows_skipped": 10.0,
                "total_rows": 10.0
            }
        },
        {
            "Name": "Rule_3",
            "Description": "Completeness \"customer_identifier\" > 0.5",
            "Result": "PASS",
            "EvaluatedMetrics": {
                "Column.customer_identifier.Completeness": 1.0
            },
            "EvaluatedRule": "Completeness \"customer_identifier\" > 0.5",
            "RuleMetrics": {
                "rows_passed": 10.0,
                "rows_failed": 0.0,
                "rows_skipped": 0.0,
                "total_rows": 10.0
            }
        }
    ],
    "AggregatedMetrics": {
        "TotalRowsProcessed": 10.0,
        "TotalRowsPassed": 10.0,
        "TotalRowsFailed": 0.0,
        "TotalRulesProcessed": 3.0,
        "TotalRulesPassed": 2.0,
        "TotalRulesFailed": 1.0
    }
```

Also, at a rule level, the following metrics are provided:

Rule Metrics:

- Rows Passed
- Rows Failed
- Row Skipped
- Total Rows Processed

2. `AggregatedMetrics` is returned as an additional data frame and `RuleOutcomes` data frame is augmented to include `RuleMetrics`.

Example implementation

The following example shows how to implement aggregated metrics in Scala:

```
// Script generated for node Evaluate Data Quality
val EvaluateDataQuality_node1741974822533_ruleset = """
  # Example rules: Completeness "colA" between 0.4 and 0.8, ColumnCount > 10
  Rules = [
      IsUnique "customer_identifier",
      RowCount > 10,
      Completeness "customer_identifier" > 0.5
  ]
"""

val EvaluateDataQuality_node1741974822533 =
 EvaluateDataQuality.processRows(frame=ChangeSchema_node1742850392012,
 ruleset=EvaluateDataQuality_node1741974822533_ruleset,
 publishingOptions=JsonOptions("""{"dataQualityEvaluationContext":
 "EvaluateDataQuality_node1741974822533", "enableDataQualityCloudWatchMetrics":
 "true", "enableDataQualityResultsPublishing": "true"}"""),
 additionalOptions=JsonOptions("""{"compositeRuleEvaluation.method":"ROW","observations.scope":
 "publishAggregatedMetrics.status": "ENABLED"}"""))

println("--------------------------------ROW LEVEL
 OUTCOMES--------------------------------")
val rowLevelOutcomes_node = EvaluateDataQuality_node1741974822533("rowLevelOutcomes")

rowLevelOutcomes_node.show(10)

 println("--------------------------------RULE LEVEL
 OUTCOMES--------------------------------")

val ruleOutcomes_node = EvaluateDataQuality_node1741974822533("ruleOutcomes")

ruleOutcomes_node.show()

 println("--------------------------------AGGREGATED
 METRICS--------------------------------")

val aggregatedMetrics_node = EvaluateDataQuality_node1741974822533("aggregatedMetrics")
```

```
aggregatedMetrics_node.show()
```

Sample results

Results are returned as follows:

```
{
    "Rule": "IsUnique \"customer_identifier\"",
    "Outcome": "Passed",
    "FailureReason": null,
    "EvaluatedMetrics": {
        "Column.customer_identifier.Uniqueness": 1
    },
    "EvaluatedRule": "IsUnique \"customer_identifier\"",
    "PassedCount": 10,
    "FailedCount": 0,
    "SkippedCount": 0,
    "TotalCount": 10
}
{
    "Rule": "RowCount > 10",
    "Outcome": "Failed",
    "FailureReason": "Value: 10 does not meet the constraint requirement!",
    "EvaluatedMetrics": {
        "Dataset.*.RowCount": 10
    },
    "EvaluatedRule": "RowCount > 10",
    "PassedCount": 0,
    "FailedCount": 0,
    "SkippedCount": 10,
    "TotalCount": 10
}
{
    "Rule": "Completeness \"customer_identifier\" > 0.5",
    "Outcome": "Passed",
    "FailureReason": null,
    "EvaluatedMetrics": {
        "Column.customer_identifier.Completeness": 1
    },
    "EvaluatedRule": "Completeness \"customer_identifier\" > 0.5",
    "PassedCount": 10,
    "FailedCount": 0,
    "SkippedCount": 0,
    "TotalCount": 10
}
```

Aggregated Metrics are as follows:

```
{ "TotalRowsProcessed": 10, "PassedRows": 10, "FailedRows": 0, "TotalRulesProcessed":
 3, "RulesPassed": 2, "RulesFailed": 1 }
```

Data Quality rule builder

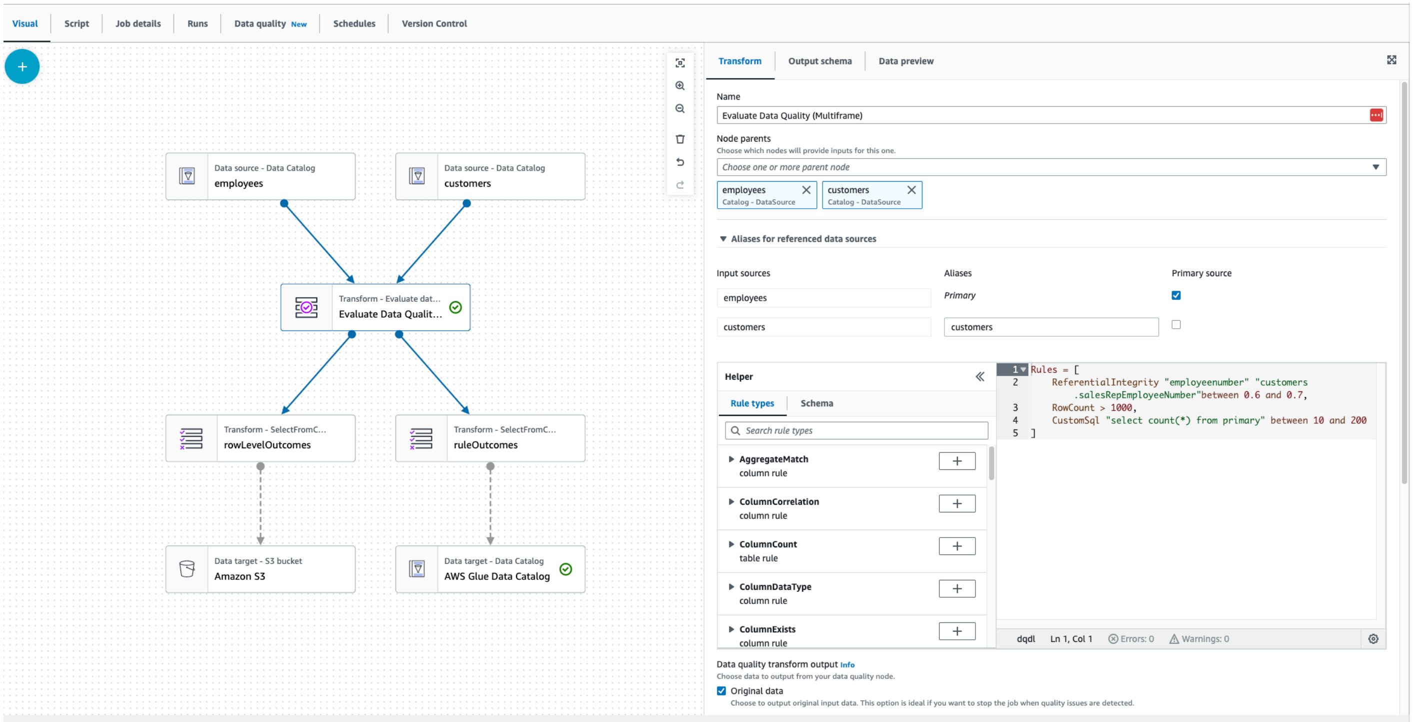
With the Data Quality Definition Language (DQDL) rule builder, you can create data quality rules to evaluate your data. Start by selecting a rule type, and then specify the parameters in the rule editor. The rule editor also shows you any errors and warnings as you create rules.

The DQDL guide provides comprehensive documentation on how to construct rules using the DQDL syntax, built-in rule types, and examples.

Evaluate Data Quality node

When you're working with the **Evaluate Data Quality** transform node and the DQDL rule builder, you can expand the working space.

- To expand the **Transform** tab to fill the entire screen, choose the expand icon in the upper-right hand corner of the node details panel.
- To expand the DQDL rule editor, choose the **<<** icon to expand the rule editor and collapse the **Rule types** and **Schema** tabs.


Visual
Script
Job details
Runs
Data quality New
Schedules
Version Control
Data source - Data Catalog
employees
Data source - Data Catalog
customers
Transform - Evaluate dat...
Evaluate Data Qualit...
Transform - SelectFromC...
rowLevelOutcomes
Transform - SelectFromC...
ruleOutcomes
Data target - S3 bucket
Amazon S3
Data target - Data Catalog
AWS Glue Data Catalog
Transform
Output schema
Data preview
Helper
Rule types
Schema
Search rule types
AggregateMatch
column rule
ColumnCorrelation
column rule
ColumnCount
table rule
ColumnDataType
column rule
ColumnExists
column rule
dqdl
Ln 1, Col 1
Errors: 0
Warnings: 0

Name: Evaluate Data Quality (Multiframe)

Node parents: Choose which nodes will provide inputs for this one.

employees (Catalog - DataSource), customers (Catalog - DataSource)

Aliases for referenced data sources

Input sources	Aliases	Primary source
employees	Primary	☑
customers	customers	☐

```
Rules = [
    ReferentialIntegrity "employeenumber" "customers
        .salesRepEmployeeNumber"between 0.6 and 0.7,
    RowCount > 1000,
    CustomSql "select count(*) from primary" between 10 and 200
]
```

Data quality transform output Info
Choose data to output from your data quality node.

☑ Original data
Choose to output original input data. This option is ideal if you want to stop the job when quality issues are detected.

Components

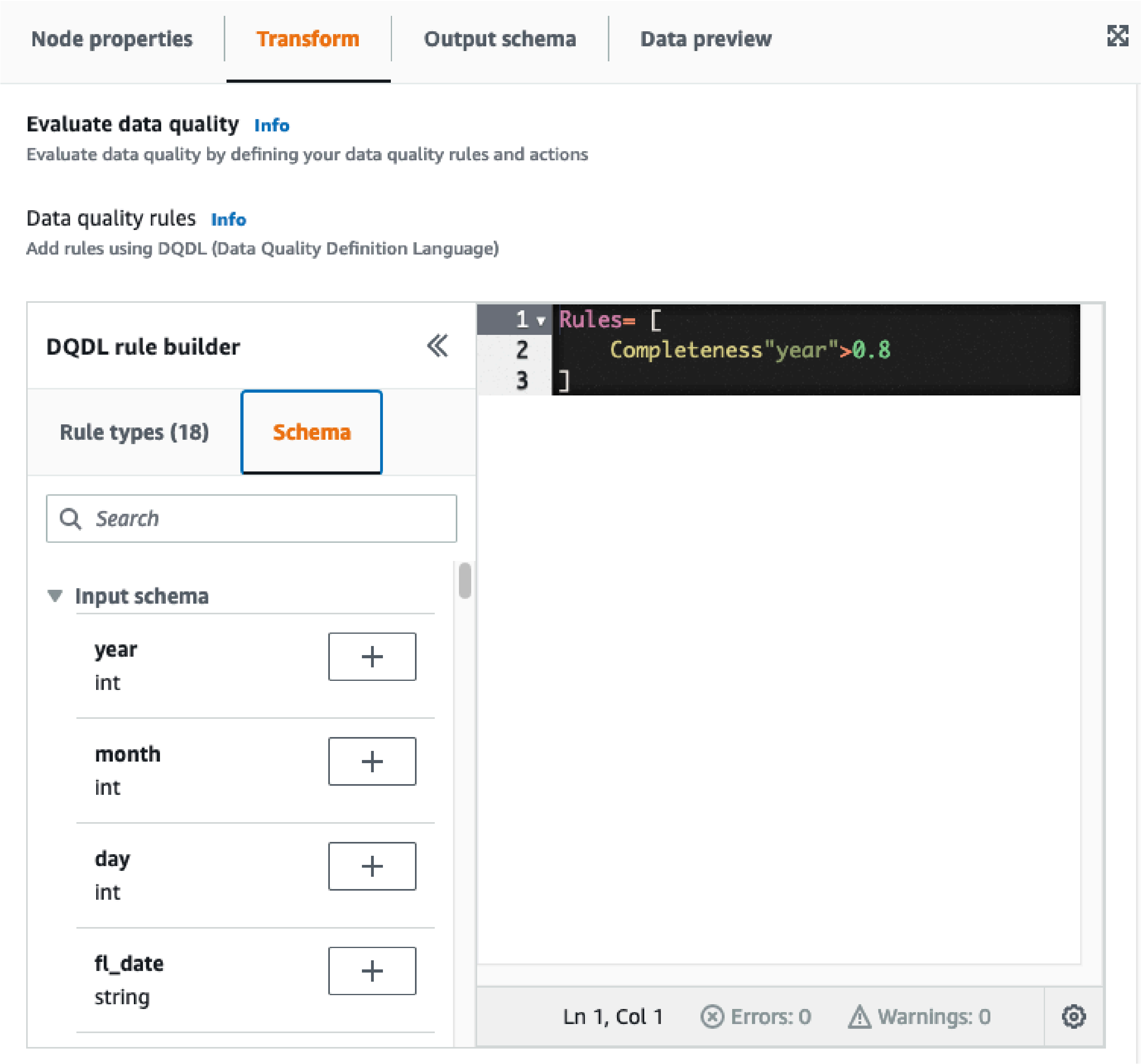
There are 26 rule types that are built into AWS Glue Studio. Each rule type has a description and examples of how they can be used.

Data quality rule types

AWS Glue Studio provides built-in rule types for ease in creating a rule. For more information on rule types, see DQDL rule type reference.

Schema

The **Schema** tab displays the column names and data type from the parent node. Schemas from multiple nodes are displayed. You can view the input schema, search by column name, and insert the column into the rule editor.


Node properties
Transform
Output schema
Data preview
Evaluate data quality
Info
Data quality rules
Info
DQDL rule builder
Rule types (18)
Schema
Search
Input schema
year
int
month
int
day
int
fl_date
string
Ln 1, Col 1
Errors: 0
Warnings: 0

Evaluate data quality by defining your data quality rules and actions

Add rules using DQDL (Data Quality Definition Language)

```
Rules= [
    Completeness"year">0.8
]
```

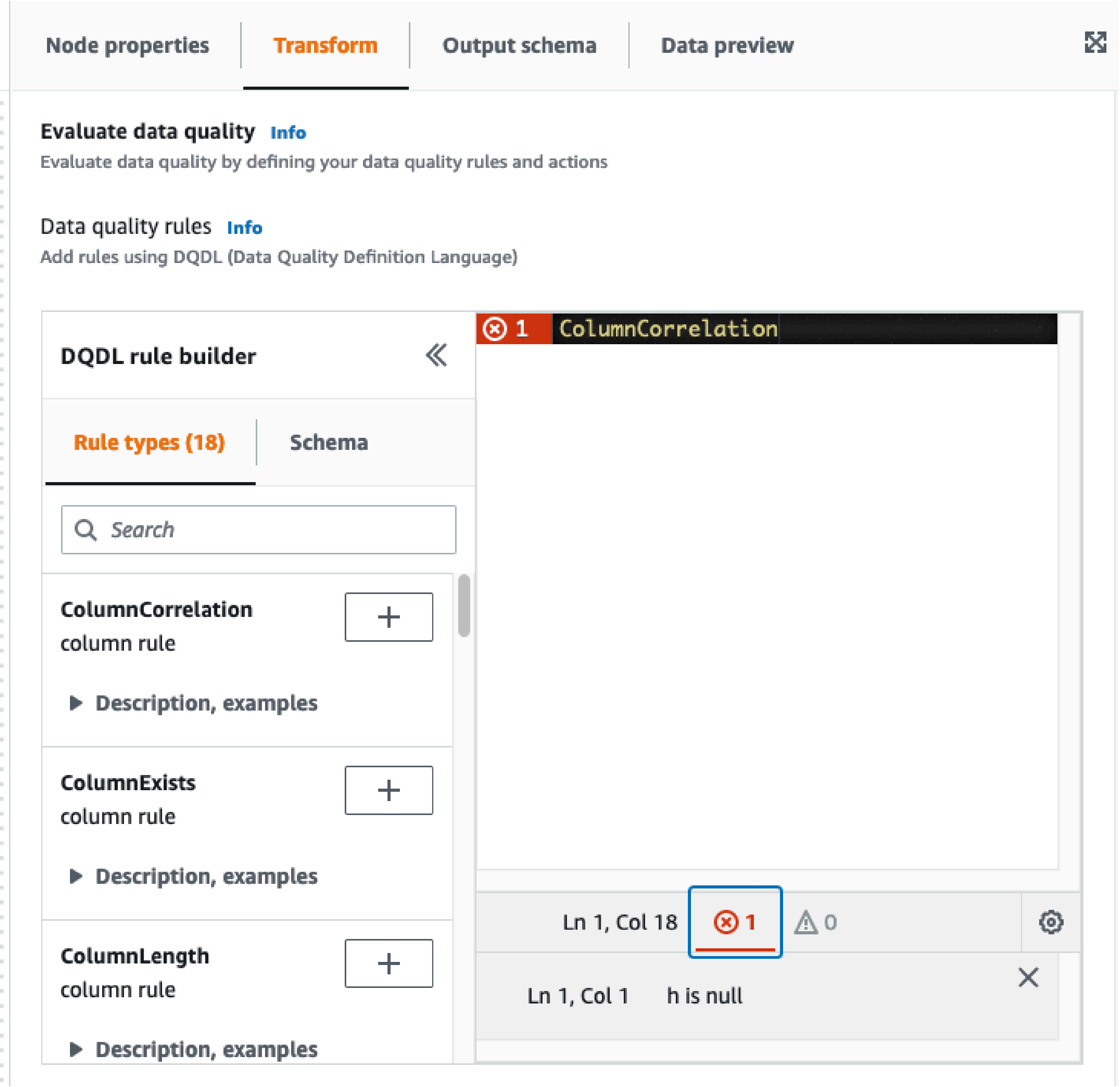
Rule editor

The rule editor is a text editor where you can write and edit rules. If you select a rule type from the DQDL rule builder, the rule type is added to the rule editor. You can then specify parameters, add rules, and edit rules as needed by modifying the text. AWS Glue Studio validates the rules in the rule editor and displays errors and warnings if there are any.

Errors and warnings

If a rule doesn't follow the DQDL rule syntax, the rule editor shows several visual indicators that there is an error:

- The rule editor displays an error icon and red color on the line with the error.
- The rule editor displays the number of errors next to the red error icon.
- When you choose the line with the error, descriptions of the error and location (line and column) are displayed at the bottom of the rule editor.


Node properties
Transform
Output schema
Data preview

Evaluate data quality Info

Evaluate data quality by defining your data quality rules and actions

Data quality rules Info

Add rules using DQDL (Data Quality Definition Language)

DQDL rule builder
Rule types (18)
Schema
Search

ColumnCorrelation column rule

▶ Description, examples

ColumnExists column rule

▶ Description, examples

ColumnLength column rule

▶ Description, examples

```
1 ColumnCorrelation
```

Ln 1, Col 18 ⊗ 1 ⚠ 0

Ln 1, Col 1 h is null

Data quality actions

By default, this action is not selected and the job will complete its run even if the data quality rules fail.

Choose between the following actions. You can use actions to publish results to CloudWatch or stop jobs based on specific criteria. Actions are only available after you create a rule.

- **Publish results to CloudWatch –** When you run a job, add the results to CloudWatch.
- **Fail job when data quality fails –** If data quality rules fail, the job will also fail as a result.

Data quality transform output

- **Original data –** Choose to output original input data. This option is ideal if you want to stop the job when quality issues are detected.
- **Data quality metrics –** Choose to output configured rules and their pass or fail status. This option is useful if you want to take a custom action.

Data quality output settings

Set the data quality result location by specifying the Amazon S3 location as the data quality output target.

Configuring anomaly detection in AWS Glue ETL jobs

To begin with anomaly detection in AWS Glue Studio, open a AWS Glue Studio job and click on the **Evaluate Data Quality Transform**.

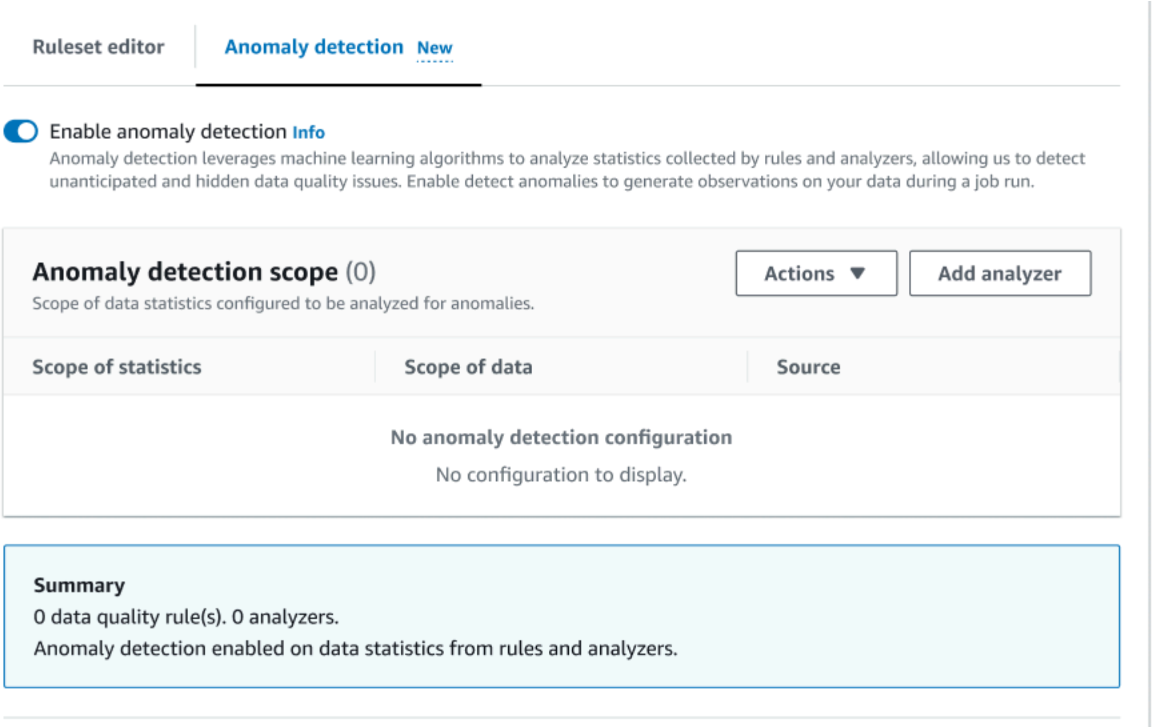
By enabling this feature, AWS Glue Data Quality will analyze your data over time to detect anomalies. It provides valuable data statistics and observations about your data, allowing you to take action on any identified anomalies.

Review the Anomaly Detection documentation to understand the inner workings of this feature.

Enabling anomaly detection

To enable anomaly detection in AWS Glue Studio:

1. Choose the **Data Quality** node in your job, then choose the **Anomaly detection** tab. Toggle to turn on **Enable Anomaly Detection**.



Enable anomaly detection Info

Anomaly detection leverages machine learning algorithms to analyze statistics collected by rules and analyzers, allowing us to detect unanticipated and hidden data quality issues. Enable detect anomalies to generate observations on your data during a job run.

Anomaly detection scope (0)

Scope of data statistics configured to be analyzed for anomalies.

Scope of statistics	Scope of data	Source

No anomaly detection configuration

No configuration to display.

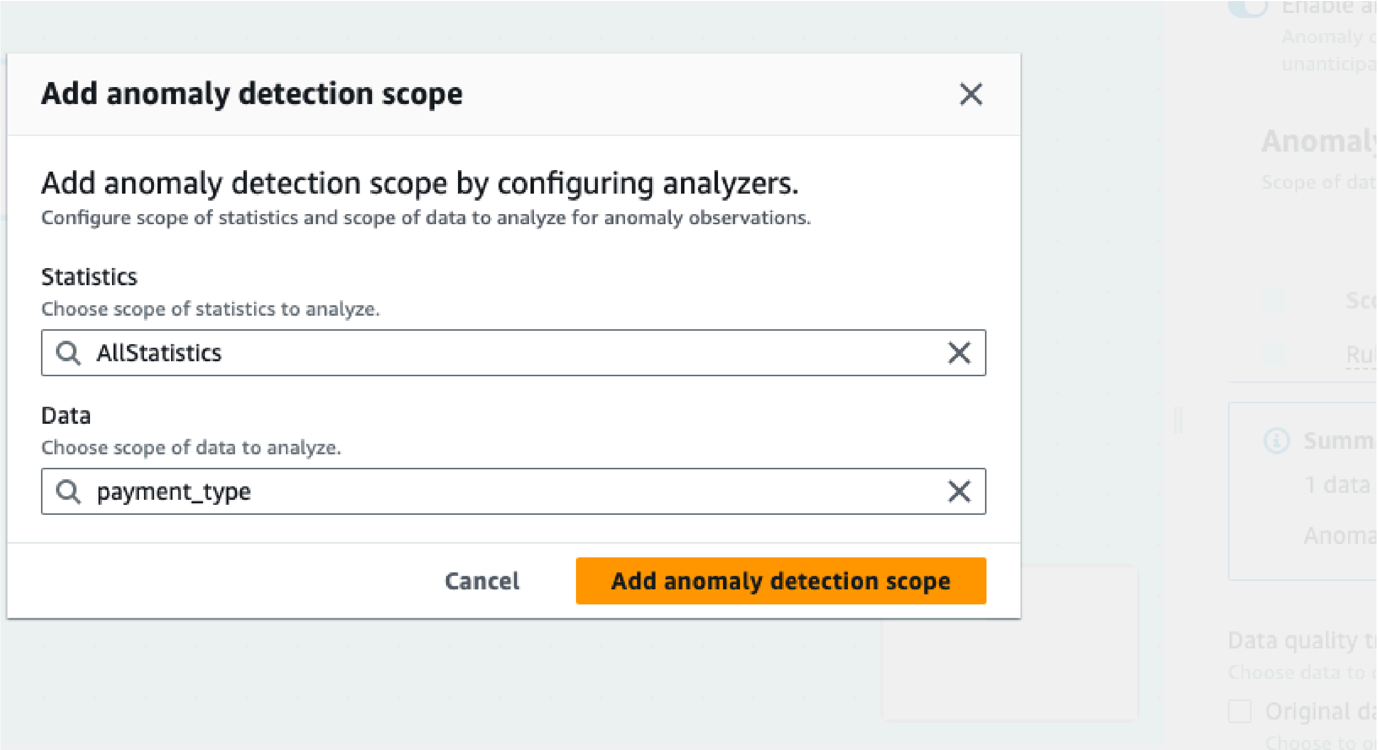
Summary

0 data quality rule(s). 0 analyzers.

Anomaly detection enabled on data statistics from rules and analyzers.

2. Define the data to monitor for anomalies by choosing **Add analyzer**. There are two fields that you can populate: Statistics and Data.

   - **Statistics** are information about your data's shape and other properties. You can choose one or more statistics at a time, or choose **All statistics**. Statistics include: Completeness, Uniqueness, Mean, Sum, StandardDeviation, Entropy, DistinctValuesCount, UniqueValueRatio and more. Refer to the Analyzers documentation for more details.
   - **Data** is the columns in your dataset. You can choose all columns or individual columns.



Add anomaly detection scope

Add anomaly detection scope by configuring analyzers.

Configure scope of statistics and scope of data to analyze for anomaly observations.

Statistics

Choose scope of statistics to analyze.

AllStatistics

Data

Choose scope of data to analyze.

payment_type

3. Choose **Add anomaly detection scope** to save your changes. After you've added analyzers, you can see them in the **Anomaly detection scope** section.

   You can also use the **Actions** menu to edit your analyzers, or choose the **Ruleset editor** tab and edit the analyzer directly in the ruleset editor notepad. You will see the analyzers that you saved under any rules that you've created.

   ```
   Rules = [

   ]

   Analyzers = [
       Completeness “id”
   ]
   ```

Once the updated ruleset and analyzers are configured, AWS Glue Data Quality continuously monitors incoming data streams. It can signal potential anomalies through alerts or job stops, depending on your settings. This proactive monitoring helps ensure data quality and integrity throughout your data pipelines.

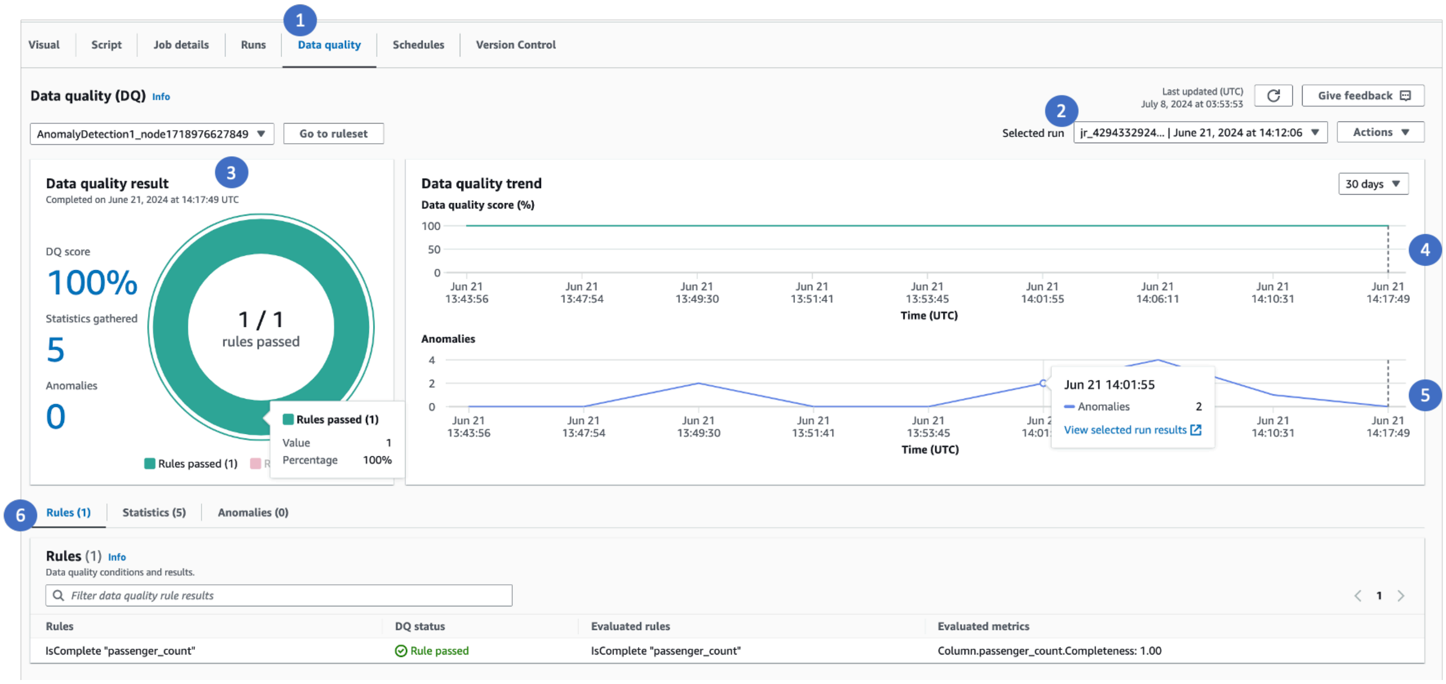
In the next section, you will learn how to effectively monitor anomalies identified by the system. You'll also learn how to view and analyze the data statistics gathered by AWS Glue Data Quality. Additionally, you'll understand how to provide feedback to the machine learning model that powers the Anomaly Detection feature. This feedback loop is crucial for improving the model's accuracy and ensuring it can effectively detect anomalies that align with your specific business requirements and data patterns.

Viewing data quality scores and anomalies

In this section, we will explore the data quality dashboard and different functionalities that it provides.

Visualize and understand high level data quality metrics and trends

Once your job is successful, choose the **Data Quality** tab to view data quality scores and anomalies.


1
2
3
4
5
6
Visual
Script
Job details
Runs
Data quality
Schedules
Version Control
Data quality (DQ)
Info
Last updated (UTC)
July 8, 2024 at 03:53:53
Give feedback
AnomalyDetection1_node1718976627849
Go to ruleset
Selected run
jr_4294332924... | June 21, 2024 at 14:12:06
Actions
Data quality result
Completed on June 21, 2024 at 14:17:49 UTC
DQ score
100%
Statistics gathered
5
Anomalies
0
1 / 1
rules passed
Rules passed (1)
Value 1
Percentage 100%
Rules passed (1)
Data quality trend
Data quality score (%)
30 days
100
50
0
Jun 21 13:43:56
Jun 21 13:47:54
Jun 21 13:49:30
Jun 21 13:51:41
Jun 21 13:53:45
Jun 21 14:01:55
Jun 21 14:06:11
Jun 21 14:10:31
Jun 21 14:17:49
Time (UTC)
Anomalies
4
2
0
Jun 21 14:01:55
Anomalies 2
View selected run results
Time (UTC)
Rules (1)
Statistics (5)
Anomalies (0)
Rules (1)
Info
Data quality conditions and results.
Filter data quality rule results
1

Rules	DQ status	Evaluated rules	Evaluated metrics
IsComplete "passenger_count"	Rule passed	IsComplete "passenger_count"	Column.passenger_count.Completeness: 1.00

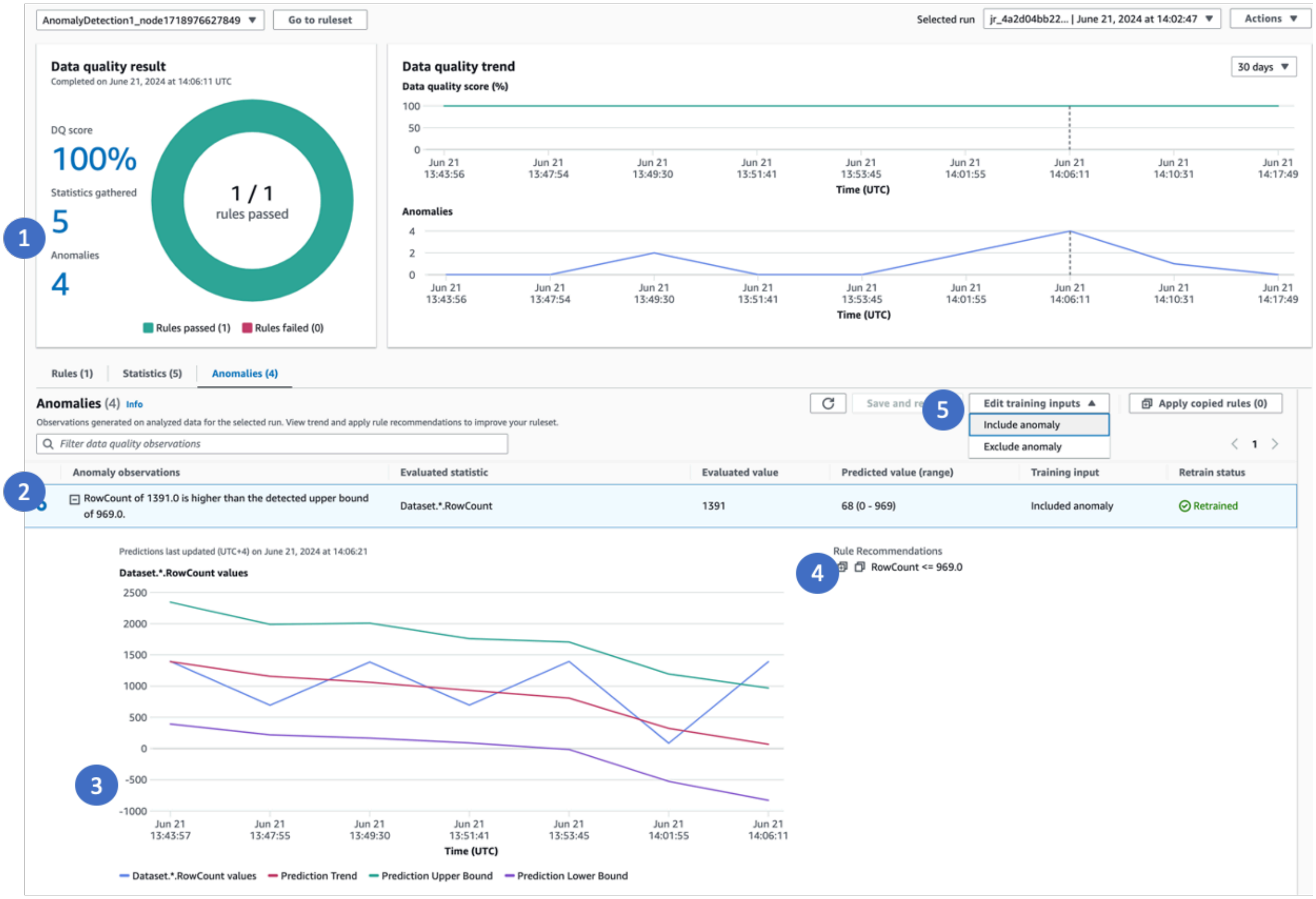
The following components in the Data quality tab provide useful information.

1. Choose the **Data Quality** tab to view data quality metrics.
2. Select a specific job run ID to view the Data Quality score.
3. This pane shows three important information. You can choose each to navigate to specific tables to view anomalies, data statistics or rules.

- Data Quality score when rules are configured.
- Number of statistics gathered by Rules and Analyzers.
- Total number of anomalies detected.

4. This trend chart shows how data quality is trending over time. You can hover over the trend and go to a specific time when data quality scores have deteriorated.
5. Anomaly trends over time will show you the number of anomalies detected over time.
6. Tabs:
   - Rules Tab is the default tab that shows list of all the rules and status. Evaluated Rules is useful in case of dynamic rules to view the actual value the rule was evaluated to.
   - Statistics Tab lists all the statistics, allowing you to view the metrics and the trends over time.
   - Anomalies tab shows the list of anomalies that were detected.

Viewing anomalies and training anomaly detection algorithm



Data quality result
Completed on June 21, 2024 at 14:06:11 UTC

DQ score 100%
Statistics gathered 5
Anomalies 4

Anomalies (4) Info
Observations generated on analyzed data for the selected run. View trend and apply rule recommendations to improve your ruleset.

Anomaly observations	Evaluated statistic	Evaluated value	Predicted value (range)	Training input	Retrain status
RowCount of 1391.0 is higher than the detected upper bound of 969.0.	Dataset.*.RowCount	1391	68 (0 - 969)	Included anomaly	Retrained

Predictions last updated (UTC+4) on June 21, 2024 at 14:06:21

Rule Recommendations
RowCount <= 969.0

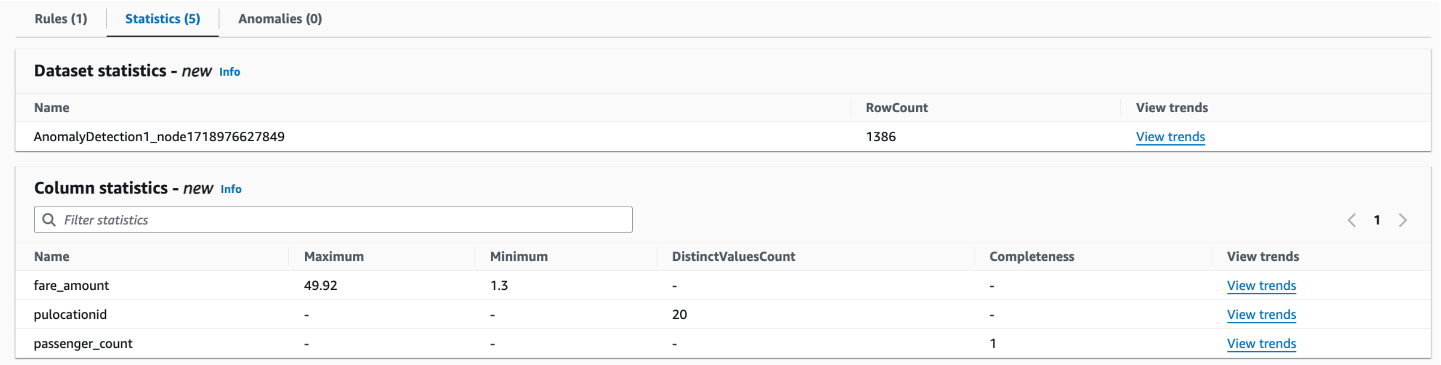
AnomalyDetection1_node1718976627849
Go to ruleset
Selected run
jr_4a2d04bb22... | June 21, 2024 at 14:02:47
Actions
1 / 1
rules passed
Rules passed (1)
Rules failed (0)
Data quality trend
Data quality score (%)
30 days
100
50
0
Jun 21 13:43:56
Jun 21 13:47:54
Jun 21 13:49:30
Jun 21 13:51:41
Jun 21 13:53:45
Jun 21 14:01:55
Jun 21 14:06:11
Jun 21 14:10:31
Jun 21 14:17:49
Time (UTC)
Anomalies
4
2
0
Rules (1)
Statistics (5)
Anomalies (4)
Save and re
Edit training inputs
Include anomaly
Exclude anomaly
Apply copied rules (0)
Filter data quality observations
1
Dataset.*.RowCount values
2500
2000
1500
1000
500
0
-500
-1000
Jun 21 13:43:57
Jun 21 13:47:55
Jun 21 13:49:30
Jun 21 13:51:41
Jun 21 13:53:45
Jun 21 14:01:55
Jun 21 14:06:11
Time (UTC)
Dataset.*.RowCount values
Prediction Trend
Prediction Upper Bound
Prediction Lower Bound
1
2
3
4
5

Call outs for the image above:

1. When anomalies are detected, click the anomaly or select the Anomalies tab
2. AWS Glue Data Quality provides a detailed explanation of the anomaly, the actual value, predicted range
3. AWS Glue Data Quality shows a trend line. It has the actual value, a derived trend based on the actual values (red line), the upper bound and the lower bound
4. AWS Glue Data Quality recommends data quality rules that can be used to capture the patterns for future. You can copy all the rules that are recommended to you and apply them to your data quality node to capture these patterns effectively.
5. You can provide inputs to the machine learning (ML) model to exclude anomalous values, ensuring that future runs detect anomalies accurately. If you don't explicitly exclude anomalies, AWS Glue Data Quality will automatically consider them as part of the model for future predictions. It's important to note that only the latest run will reflect the model inputs you provide. For example, if you went back and excluded anomalous points from a few previous runs, the model will not reflect those changes unless you view and update the model inputs in the latest run. The model will continue to use the previously provided inputs until you make necessary adjustments in the most recent run. By actively managing the exclusion of anomalous values, you can refine the ML model's understanding of what constitutes an anomaly for your specific data patterns and requirements, leading to more accurate anomaly detection over time.

Viewing Data Statistics over time and providing training inputs

Sometimes, you may want to view data statistics or data profiles and view how they are progressing over time. To do this, choose **Statistics** or open the **Statistics** tab. You can then view the latest data statistics gathered by AWS Glue Data Quality.



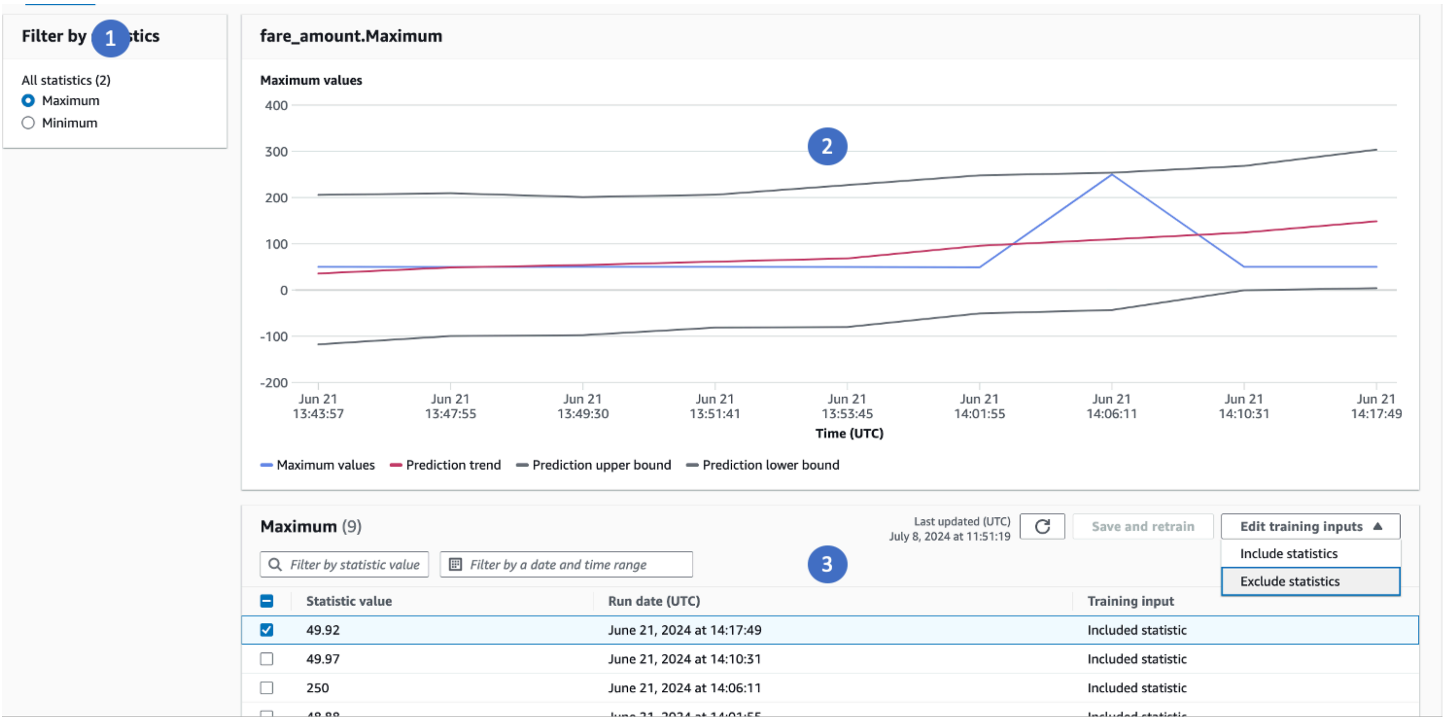
Dataset statistics - *new* Info

Name	RowCount	View trends
AnomalyDetection1_node1718976627849	1386	View trends

Column statistics - *new* Info

Name	Maximum	Minimum	DistinctValuesCount	Completeness	View trends
fare_amount	49.92	1.3	-	-	View trends
pulocationid	-	-	20	-	View trends
passenger_count	-	-	-	1	View trends

Clicking **View Trends** shows you how each of the statistics are progressing over time.


Filter by statistics
1
All statistics (2)
● Maximum
○ Minimum
fare_amount.Maximum
Maximum values
400
300
200
100
0
-100
-200
2
Jun 21 13:43:57
Jun 21 13:47:55
Jun 21 13:49:30
Jun 21 13:51:41
Jun 21 13:53:45
Jun 21 14:01:55
Jun 21 14:06:11
Jun 21 14:10:31
Jun 21 14:17:49
Time (UTC)
Maximum values
Prediction trend
Prediction upper bound
Prediction lower bound
Maximum (9)
Last updated (UTC) July 8, 2024 at 11:51:19
Save and retrain
Edit training inputs ▲
Include statistics
Exclude statistics
Filter by statistic value
Filter by a date and time range
3

☐	Statistic value	Run date (UTC)	Training input
☑	49.92	June 21, 2024 at 14:17:49	Included statistic
☐	49.97	June 21, 2024 at 14:10:31	Included statistic
☐	250	June 21, 2024 at 14:06:11	Included statistic

1. You can select the statistic for a specified column
2. You can view how the trends are progressing
3. You can select anomalous values and choose to exclude or include them. By providing this feedback, the algorithm will either exclude or include the identified anomalous data points and retrain the model. This retraining process ensures accurate anomaly detection moving forward, as the model learns from the feedback you've provided about which values should be considered anomalous or not.

   Through this feedback loop, you have the ability to refine the algorithm's understanding of what constitutes an anomaly for your specific data patterns and business requirements. By excluding values that should not be flagged as anomalies, or including values that were missed, the retrained model will become better at differentiating between expected and truly anomalous data points.

Data Quality for ETL jobs in AWS Glue Studio notebooks

In this tutorial, you learn how to use AWS Glue Data Quality for extract, transform, and load (ETL) jobs in AWS Glue Studio notebooks.

You can use notebooks in AWS Glue Studio to edit job scripts and view the output without having to run a full job. You can also add markdown and save notebooks as `.ipynb` files and job scripts. Note that you can start a notebook without installing software locally or managing servers. When you're satisfied with your code, you can use AWS Glue Studio to easily convert your notebook to an AWS Glue job.

The dataset that you use in this example consists of Medicare Provider payment data that was downloaded from two *Data.CMS.gov* datasets: "Inpatient Prospective Payment System Provider Summary for the Top 100 Diagnosis-Related Groups - FY2011" and "Inpatient Charge Data FY 2011".

After downloading the data, we modified the dataset to introduce a couple of erroneous records at the end of the file. This modified file is located in a public Amazon S3 bucket at `s3://awsglue-datasets/examples/medicare/Medicare_Hospital_Provider.csv`.

Prerequisites

- AWS Glue role with Amazon S3 permission to write to your destination Amazon S3 bucket
- A new notebook (see Getting started with notebooks in AWS Glue Studio)

Creating an ETL job in AWS Glue Studio

To create an ETL job

1. Change the session version to AWS Glue 3.0.

   To do this, remove all boilerplate code cells with the following magic and run the cell. Note that this boilerplate code is automatically provided in the first cell when a new notebook is created.

   ```
   %glue_version 3.0
   ```

2. Copy the following code and run it in the cell.

   ```
   import sys
   from awsglue.transforms import *
   from awsglue.utils import getResolvedOptions
   from pyspark.context import SparkContext
   from awsglue.context import GlueContext
   ```

```
from awsglue.job import Job

sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
```

3. In the next cell, import the `EvaluateDataQuality` class that evaluatesAWS Glue Data Quality.

```
from awsgluedq.transforms import EvaluateDataQuality
```

4. In the next cell, read in the source data by using the .csv file that's stored in the public Amazon S3 bucket.

```
medicare = spark.read.format(
"csv").option(
"header", "true").option(
"inferSchema", "true").load(
's3://awsglue-datasets/examples/medicare/Medicare_Hospital_Provider.csv')
medicare.printSchema()
```

5. Convert the data to an AWS Glue DynamicFrame.

```
from awsglue.dynamicframe import DynamicFrame
medicare_dyf = DynamicFrame.fromDF(medicare,glueContext,"medicare_dyf")
```

6. Create the ruleset using Data Quality Definition Language (DQDL).

```
EvaluateDataQuality_ruleset = """
    Rules = [
        ColumnExists "Provider Id",
        IsComplete "Provider Id",
        ColumnValues  " Total Discharges " > 15
]
    ]
"""
```

7. Validate the dataset against the ruleset.

```
EvaluateDataQualityMultiframe = EvaluateDataQuality().process_rows(
    frame=medicare_dyf,
    ruleset=EvaluateDataQuality_ruleset,
    publishing_options={
        "dataQualityEvaluationContext": "EvaluateDataQualityMultiframe",
        "enableDataQualityCloudWatchMetrics": False,
        "enableDataQualityResultsPublishing": False,
    },
    additional_options={"performanceTuning.caching": "CACHE_NOTHING"},
)
```

8. Review the results.

```
ruleOutcomes = SelectFromCollection.apply(
    dfc=EvaluateDataQualityMultiframe,
    key="ruleOutcomes",
    transformation_ctx="ruleOutcomes",
)

ruleOutcomes.toDF().show(truncate=False)
```

Output:

```
--------------------------------------+-------
+-----------------------------------------------------
+-------------------------------------------+
|Rule                                  |Outcome|FailureReason
                  |EvaluatedMetrics                           |
+--------------------------------------+-------
+-----------------------------------------------------
+-------------------------------------------+
|ColumnExists "Provider Id"            |Passed |null
                  |{}                                         |
|IsComplete "Provider Id"              |Passed |null
                  |{Column.Provider Id.Completeness -> 1.0}   |
|ColumnValues " Total Discharges " > 15|Failed |Value: 11.0 does not meet the
 constraint requirement!|{Column. Total Discharges .Minimum -> 11.0}|
+--------------------------------------+-------
+-----------------------------------------------------
+-------------------------------------------+
```

9. Filter passed rows and review the failed rows from the Data Quality row-level results.

```
owLevelOutcomes = SelectFromCollection.apply(
dfc=EvaluateDataQualityMultiframe,
key="rowLevelOutcomes",
transformation_ctx="rowLevelOutcomes",
)

rowLevelOutcomes_df = rowLevelOutcomes.toDF() # Convert Glue DynamicFrame to
 SparkSQL DataFrame
rowLevelOutcomes_df_passed =
 rowLevelOutcomes_df.filter(rowLevelOutcomes_df.DataQualityEvaluationResult ==
 "Passed") # Filter only the Passed records.
rowLevelOutcomes_df.filter(rowLevelOutcomes_df.DataQualityEvaluationResult ==
 "Failed").show(5, truncate=False) # Review the Failed records
```

Output:

```
+----------------------------------------+-----------
+-------------------------------------+--------------------------+-------------
+--------------+-----------------+------------------------------------
+------------------+-------------------------+------------------------
+-------------------------+--------------------------
+----------------------------------------+----------------------------
+---------------------------+
|DRG Definition                          |Provider Id|Provider Name
           |Provider Street Address   |Provider City|Provider State|Provider Zip
 Code|Hospital Referral Region Description| Total Discharges | Average Covered
 Charges | Average Total Payments |Average Medicare Payments|DataQualityRulesPass
      |DataQualityRulesFail                    |DataQualityRulesSkip        |
DataQualityEvaluationResult|
+----------------------------------------+-----------
+-------------------------------------+--------------------------+-------------
+--------------+-----------------+------------------------------------
+------------------+-------------------------+------------------------
+-------------------------+--------------------------
```

```
+----------------------------------------+----------------------------
+---------------------------+
|039 - EXTRACRANIAL PROCEDURES W/O CC/MCC|10005      |MARSHALL MEDICAL CENTER SOUTH
        |2505 U S HIGHWAY 431 NORTH|BOAZ         |AL            |35957
 |AL - Birmingham                     |14               |$15131.85
  |$5787.57                |$4976.71                 |[IsComplete "Provider Id"]|
[ColumnValues " Total Discharges " > 15]|[ColumnExists "Provider Id"]|Failed
              |
|039 - EXTRACRANIAL PROCEDURES W/O CC/MCC|10046      |RIVERVIEW REGIONAL MEDICAL
 CENTER    |600 SOUTH THIRD STREET    |GADSDEN      |AL            |35901
   |AL - Birmingham                     |14               |$67327.92
   |$5461.57                |$4493.57                 |[IsComplete "Provider Id"]|
[ColumnValues " Total Discharges " > 15]|[ColumnExists "Provider Id"]|Failed
              |
|039 - EXTRACRANIAL PROCEDURES W/O CC/MCC|10083      |SOUTH BALDWIN REGIONAL
 MEDICAL CENTER|1613 NORTH MCKENZIE STREET|FOLEY        |AL            |36535
        |AL - Mobile                         |15               |$25411.33
         |$5282.93                |$4383.73                 |[IsComplete "Provider
 Id"]|[ColumnValues " Total Discharges " > 15]|[ColumnExists "Provider Id"]|Failed
                    |
|039 - EXTRACRANIAL PROCEDURES W/O CC/MCC|30002      |BANNER GOOD SAMARITAN MEDICAL
 CENTER |1111 EAST MCDOWELL ROAD   |PHOENIX      |AZ            |85006
 |AZ - Phoenix                        |11               |$34803.81
  |$7768.90                |$6951.45                 |[IsComplete "Provider Id"]|
[ColumnValues " Total Discharges " > 15]|[ColumnExists "Provider Id"]|Failed
              |
|039 - EXTRACRANIAL PROCEDURES W/O CC/MCC|30010      |CARONDELET ST  MARYS HOSPITAL
        |1601 WEST ST MARY'S ROAD  |TUCSON       |AZ            |85745
 |AZ - Tucson                         |12               |$35968.50
  |$6506.50                |$5379.83                 |[IsComplete "Provider Id"]|
[ColumnValues " Total Discharges " > 15]|[ColumnExists "Provider Id"]|Failed
              |
+----------------------------------------+-----------
+-------------------------------------+--------------------------+-------------
+--------------+-----------------+------------------------------------
+------------------+-------------------------+------------------------
+-------------------------+--------------------------
+----------------------------------------+----------------------------
+---------------------------+
only showing top 5 rows
```

Note that AWS Glue Data Quality added four new columns (DataQualityRulesPass, DataQualityRulesFail, DataQualityRulesSkip, and DataQualityEvaluationResult). This indicates the records that passed, the records that failed, rules skipped for row-level evaluation, and the overall row-level results.

10. Write the output to an Amazon S3 bucket to analyze the data and visualize the results.

```
#Write the Passed records to the destination.

glueContext.write_dynamic_frame.from_options(
       frame = rowLevelOutcomes_df_passed,
       connection_type = "s3",
       connection_options = {"path": "s3://glue-sample-target/output-dir/
medicare_parquet"},
       format = "parquet")
```

Data Quality Definition Language (DQDL) reference

Data Quality Definition Language (DQDL) is a domain specific language that you use to define rules for AWS Glue Data Quality.

This guide introduces key DQDL concepts to help you understand the language. It also provides a reference for DQDL rule types with syntax and examples. Before you use this guide, we recommend that you have familiarity with AWS Glue Data Quality. For more information, see AWS Glue Data Quality.

Note

DynamicRules are only supported in AWS Glue ETL.

Contents

- DQDL syntax
  - Rule structure
  - Composite rules
    - How Composite rules work

  - Expressions
    - Keywords for NULL, EMPTY and WHITESPACES_ONLY
    - Filtering with Where Clause
  - Dynamic rules
  - Analyzers
  - Comments
- DQDL rule type reference
  - AggregateMatch
  - ColumnCorrelation
  - ColumnCount
  - ColumnDataType
  - ColumnExists
  - ColumnLength
  - ColumnNamesMatchPattern
  - ColumnValues
  - Completeness
  - CustomSQL
  - DataFreshness
  - DatasetMatch
  - DistinctValuesCount
  - Entropy
  - IsComplete
  - IsPrimaryKey
  - IsUnique
  - Mean
  - ReferentialIntegrity
  - RowCount
  - RowCountMatch
  - StandardDeviation
  - Sum

- SchemaMatch
- Uniqueness
- UniqueValueRatio
- DetectAnomalies
- FileFreshness
- FileMatch
- FileUniqueness
- FileSize

DQDL syntax

A DQDL document is case sensitive and contains a *ruleset*, which groups individual data quality rules together. To construct a ruleset, you must create a list named `Rules` (capitalized), delimited by a pair of square brackets. The list should contain one or more comma-separated DQDL rules like the following example.

```
Rules = [
   IsComplete "order-id",
   IsUnique "order-id"
]
```

Rule structure

The structure of a DQDL rule depends on the rule type. However, DQDL rules generally fit the following format.

```
<RuleType> <Parameter> <Parameter> <Expression>
```

`RuleType` is the case-sensitive name of the rule type that you want to configure. For example, `IsComplete`, `IsUnique`, or `CustomSql`. Rule parameters differ for each rule type. For a complete reference of DQDL rule types and their parameters, see DQDL rule type reference.

Composite rules

DQDL supports the following logical operators that you can use to combine rules. These rules are called Composite Rules.

and

The logical `and` operator results in `true` if and only if the rules that it connects are `true`. Otherwise, the combined rule results in `false`. Each rule that you connect with the `and` operator must be surrounded by parentheses.

The following example uses the `and` operator to combine two DQDL rules.

```
(IsComplete "id") and (IsUnique "id")
```

or

The logical `or` operator results in `true` if and only if one or more of the rules that it connects are `true`. Each rule that you connect with the `or` operator must be surrounded by parentheses.

The following example uses the `or` operator to combine two DQDL rules.

```
(RowCount "id" > 100) or (IsPrimaryKey "id")
```

You can use the same operator to connect multiple rules, so the following rule combination is allowed.

```
(Mean "Star_Rating" > 3) and (Mean "Order_Total" > 500) and (IsComplete "Order_Id")
```

You can combine the logical operators into a single expression. For example:

```
(Mean "Star_Rating" > 3) and ((Mean "Order_Total" > 500) or (IsComplete "Order_Id"))
```

You can also author more complex, nested rules.

```
(RowCount > 0) or ((IsComplete "colA") and (IsUnique "colA"))
```

How Composite rules work

By default, Composite Rules are evaluated as individual rules across the entire dataset or table and then the results are combined. In other words, it evaluates the entire column first and then applies the operator. This default behaviour is explained below with an example:

```
# Dataset
```

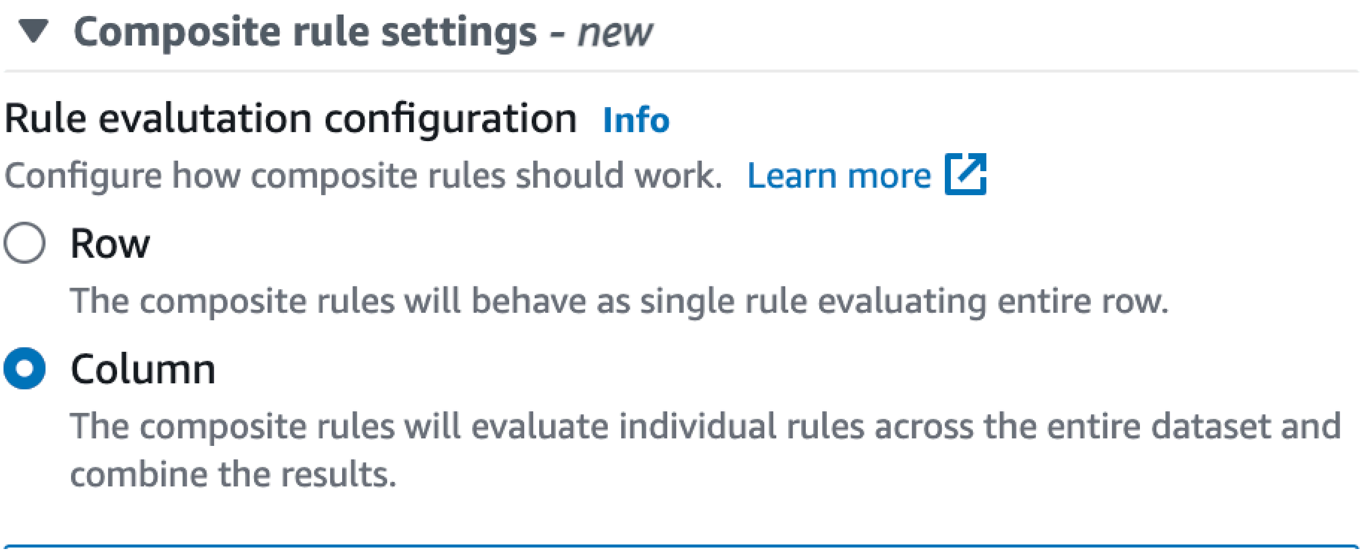
```
+------+------+
|myCol1|myCol2|
+------+------+
|     2|     1|
|     0|     3|
+------+------+

# Overall outcome

+----------------------------------------------------------+-------+
|Rule                                                      |Outcome|
+----------------------------------------------------------+-------+
|(ColumnValues "myCol1" > 1) OR (ColumnValues "myCol2" > 2)|Failed |
+----------------------------------------------------------+-------+
```

In the above example, AWS Glue Data Quality first evaluates `(ColumnValues "myCol1" > 1)` which will result in a failure. Then it will evaluate `(ColumnValues "myCol2" > 2)` which will also fail. The combination of both results will be noted as FAILED.

However, if you prefer an SQL like behaviour, where you need the entire row to be evaluated, you have to explicitly set the `ruleEvaluation.scope` parameter as shown in `additionalOptions` in the code snippet below.

```
object GlueApp {
  val datasource = glueContext.getCatalogSource(
    database="<db>",
    tableName="<table>",
    transformationContext="datasource"
  ).getDynamicFrame()

  val ruleset = """
    Rules = [
        (ColumnValues "age" >= 26) OR (ColumnLength "name" >= 4)
    ]
  """

  val dq_results = EvaluateDataQuality.processRows(
    frame=datasource,
    ruleset=ruleset,
    additionalOptions=JsonOptions("""
        {
            "compositeRuleEvaluation.method":"ROW"
          }
        """
      )
    )
}
```

In AWS Glue Data Catalog, you can easily configure this option in the user interface as shown below.



▼ Composite rule settings - *new*

Rule evalutation configuration **Info**

Configure how composite rules should work. Learn more

○ Row

The composite rules will behave as single rule evaluating entire row.

● Column

The composite rules will evaluate individual rules across the entire dataset and combine the results.

Once set, the composite rules will behave as a single rule evaluating the entire row. The following example illustrates this behaviour.

```


# Row Level outcome

+------+------+------------------------------------------------------------
+---------------------------+
|myCol1|myCol2|DataQualityRulesPass                                        |
DataQualityEvaluationResult|
+------+------+------------------------------------------------------------
+---------------------------+
|2     |1     |[(ColumnValues "myCol1" > 1) OR (ColumnValues "myCol2" > 2)]|Passed
                 |
|0     |3     |[(ColumnValues "myCol1" > 1) OR (ColumnValues "myCol2" > 2)]|Passed
                 |
+------+------+------------------------------------------------------------
+---------------------------+
```

Some rules cannot be supported in this feature because their overall outcome rely on thresholds or ratios. They are listed below.

Rules relying on ratios:

- Completeness
- DatasetMatch
- ReferentialIntegrity
- Uniqueness

Rules dependent on thresholds:

When the following rules include with threshold, they are not supported. However, rules that do not involve `with threshold` remain supported.

- ColumnDataType
- ColumnValues
- CustomSQL

Expressions

If a rule type doesn't produce a Boolean response, you must provide an expression as a parameter in order to create a Boolean response. For example, the following rule checks the mean (average) of all the values in a column against an expression to return a true or false result.

```
Mean "colA" between 80 and 100
```

Some rule types such as `IsUnique` and `IsComplete` already return a Boolean response.

The following table lists expressions that you can use in DQDL rules.

Supported DQDL expressions

Expression	Description	Example
=*x*	Resolves to `true` if the rule type response is equal to *x*.	`Completeness "colA" = "1.0",` `ColumnValues "colA" = "2022-06-30"`
!=*x*	x Resolves to true if the rule type response is not equal to *x*.	`ColumnValues "colA" != "a",` `ColumnValues "colA" != "2022-06-30"`
> *x*	Resolves to `true` if the rule type response is greater than *x*.	`ColumnValues "colA" > 10`
< *x*	Resolves to `true` if the rule type response is less than *x*.	`ColumnValues "colA" < 1000,` `ColumnValues "colA" < "2022-06-30"`
>= *x*	Resolves to `true` if the rule type response is greater than or equal to *x*.	`ColumnValues "colA" >= 10`
<= *x*	Resolves to `true` if the rule type response is less than or equal to *x*.	`ColumnValues "colA" <= 1000`
between *x* and *y*	Resolves to `true` if the rule type response falls in a specified range (exclusive). Only use this expression type for numeric and date types.	`Mean "colA" between 8 and 100,` `ColumnValues "colA" between "2022-05-31" and "2022-06-30"`
not between *x* and *y*	Resolves to true if the rule type response does not fall in a specified range (inclusiv	`ColumnValues "colA" not between "2022-05-31" and "2022-06-30"`

Expression	Description	Example
	e). You should only use this expression type for numeric and date types.	
in [*a, b, c, ...*]	Resolves to `true` if the rule type response is in the specified set.	`ColumnValues "colA" in [ 1, 2, 3 ],` `ColumnValues "colA" in [ "a", "b", "c" ]`
not in [*a, b, c, ...*]	Resolves to `true` if the rule type response is not in the specified set.	`ColumnValues "colA" not in [ 1, 2, 3 ],` `ColumnValues "colA" not in [ "a", "b", "c" ]`
matches */ab+c/i*	Resolves to `true` if the rule type response matches a regular expression.	`ColumnValues "colA" matches "[a-zA-Z]*"`
not matches */ab+c/i*	Resolves to `true` if the rule type response does not match a regular expression.	`ColumnValues "colA" not matches "[a-zA-Z]*"`
`now()`	Works only with the `ColumnValues` rule type to create a date expression.	`ColumnValues "load_date" > (now() - 3 days)`

matches/in […]/not matches/not in […] `with threshold`	Specifies the percentage of values that match the rule conditions. Works only with the `ColumnValues`, `ColumnDataType`, and `CustomSQL` rule types.	`ColumnValues "colA" in ["A", "B"] with threshold > 0.8,`
		`ColumnValues "colA" matches "[a-zA-Z]*" with threshold between 0.2 and 0.9`
		`ColumnDataType "colA" = "Timestamp" with threshold > 0.9`

Keywords for NULL, EMPTY and WHITESPACES_ONLY

If you want to validate if a string column has a null, empty or a string with only whitespaces you can use the following keywords:

- NULL / null – This keyword resolves to true for a `null` value in a string column.

  `ColumnValues "colA" != NULL with threshold > 0.5` would return true if more than 50% of your data does not have null values.

  `(ColumnValues "colA" = NULL) or (ColumnLength "colA" > 5)` would return true for all rows which either have a null value or have length >5. *Note that this will require the use of the “compositeRuleEvaluation.method” = “ROW” option.*
- EMPTY / empty – This keyword resolves to true for an empty string (“”) value in a string column. Some data formats transform nulls in a string column to empty strings. This keyword helps filter out empty strings in your data.

  `(ColumnValues "colA" = EMPTY) or (ColumnValues "colA" in ["a", "b"])` would return true if a row is either empty, “a” or “b”. *Note that this requires the use of the “compositeRuleEvaluation.method” = “ROW” option.*
- WHITESPACES_ONLY / whitespaces_only – This keyword resolves to true for a string with only whitespaces (“ ”) value in a string column.

  `ColumnValues "colA" not in ["a", "b", WHITESPACES_ONLY]` would return true if a row is neither “a” or “b” nor just whitespaces.

  Supported rules:

  - ColumnValues

For a numeric or date based expression, if you want to validate if a column has a null you can use the following keywords.

- NULL / null – This keyword resolves to true for a null value in a string column.

  `ColumnValues "colA" in [NULL, "2023-01-01"]` would return true if a dates in your column are either `2023-01-01` or null.

`(ColumnValues "colA" = NULL) or (ColumnValues "colA" between 1 and 9)` would return true for all rows which either have a null value or have values between 1 and 9. *Note that this will require the use of the "compositeRuleEvaluation.method" = "ROW" option.*

Supported rules:

- ColumnValues

Filtering with Where Clause

Note

Where Clause is only supported in AWS Glue 4.0.

You can filter your data when authoring rules. This is helpful when you want to apply conditional rules.

```
<DQDL Rule> where "<valid SparkSQL where clause> "
```

The filter must be specified with the `where` keyword, followed by a valid SparkSQL statement that is enclosed in quotes `("")`.

If the rule you wish to add the where clause to a rule with a threshold, the where clause should be specified before the threshold condition.

```
<DQDL Rule> where "valid SparkSQL statement>" with threshold <threshold condition>
```

With this syntax you can write rules like the following.

```
Completeness "colA" > 0.5 where "colB = 10"
ColumnValues "colB" in ["A", "B"] where "colC is not null" with threshold > 0.9
ColumnLength "colC" > 10 where "colD != Concat(colE, colF)"
```

We will validate that the SparkSQL statement provided is valid. If invalid, the rule evaluation will fail and we will throw the an `IllegalArgumentException` with the following format:

```
Rule <DQDL Rule> where "<invalid SparkSQL>" has provided an invalid where clause :
```

```
<SparkSQL Error>
```

Where clause behaviour when Row level error record identification is turned on

With AWS Glue Data Quality, you can identify specific records that failed. When applying a where clause to rules that support row level results, we will label the rows that are filtered out by the where clause as `Passed`.

If you prefer to separately label the filtered out rows as `SKIPPED`, you can set the following `additionalOptions` for the ETL job.

```
object GlueApp {
  val datasource = glueContext.getCatalogSource(
    database="<db>",
    tableName="<table>",
    transformationContext="datasource"
  ).getDynamicFrame()

  val ruleset = """
    Rules = [
       IsComplete "att2" where "att1 = 'a'"
    ]
  """

  val dq_results = EvaluateDataQuality.processRows(
    frame=datasource,
    ruleset=ruleset,
    additionalOptions=JsonOptions("""
        {
          "rowLevelConfiguration.filteredRowLabel":"SKIPPED"
        }
      """
    )
  )
}
```

As an example, refer to the following rule and dataframe:

```
IsComplete att2 where "att1 = 'a'"
```

id	att1	att2	Row-level Results (Default)	Row Level Results (Skipped Option)	Comments
1	a	f	PASSED	PASSED	
2	b	d	PASSED	SKIPPED	Row is filtered out, since `att1` is not `"a"`
3	a	null	FAILED	FAILED	
4	a	f	PASSED	PASSED	
5	b	null	PASSED	SKIPPED	Row is filtered out, since `att1` is not `"a"`
6	a	f	PASSED	PASSED	

Dynamic rules

> **Note**
>
> Dynamic Rules are only supported in AWS Glue ETL and is not supported in AWS Glue Data Catalog.

You can now author dynamic rules to compare current metrics produced by your rules with their historical values. These historical comparisons are enabled by using the `last()` operator in expressions. For example, the rule `RowCount > last()` will succeed when the number of rows in the current run is greater than the most recent prior row count for the same dataset. `last()` takes an optional natural number argument describing how many prior metrics to consider; `last(k)` where `k >= 1` will reference the last k metrics.

- If no data points are available, `last(k)` will return the default value 0.0.
- If fewer than `k` metrics are available, `last(k)` will return all prior metrics.

To form valid expressions use `last(k)`, where `k > 1` requires an aggregation function to reduce multiple historical results to a single number. For example, `RowCount > avg(last(5))` will check whether the current dataset's row count is strictly greater than the average of the last five row counts for the same dataset. `RowCount > last(5)` will produce an error because the current dataset row count can't be meaningfully compared to a list.

Supported aggregation functions:

- `avg`
- `median`
- `max`
- `min`
- `sum`
- `std` (standard deviation)
- `abs` (absolute value)
- `index(last(k), i)` will allow for selecting the `i`th most recent value out of the last `k`. `i` is zero-indexed, so `index(last(3), 0)` will return the most recent datapoint and `index(last(3), 3)` will result in an error as there are only three datapoints and we attempt to index the 4th most recent one.

Sample expressions

ColumnCorrelation

- `ColumnCorrelation "colA" "colB" < avg(last(10))`

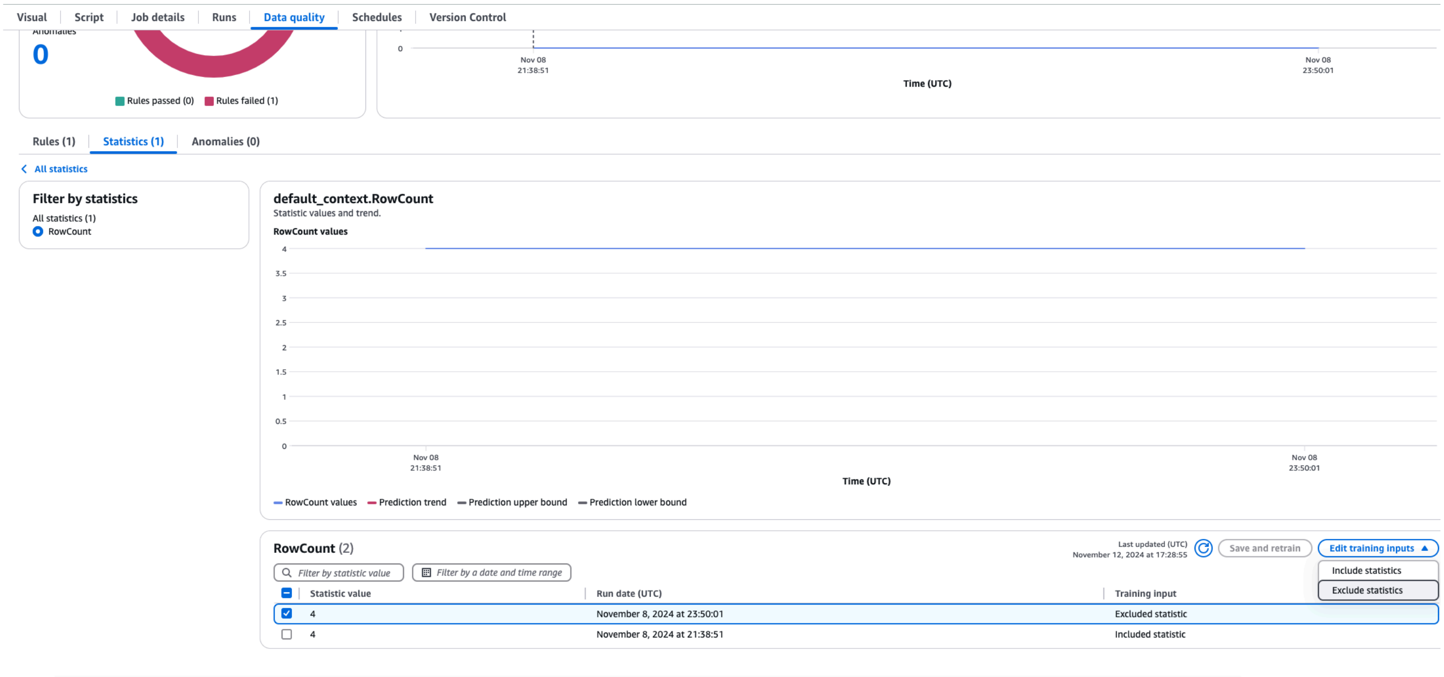
DistinctValuesCount

- `DistinctValuesCount "colA" between min(last(10))-1 and max(last(10))+1`

Most rule types with numeric conditions or thresholds support dynamic rules; see the provided table, Analyzers and Rules, to determine whether dynamic rules are supported for your rule type.

Excluding statistics from dynamic rules

Sometimes, you will need to exclude data statistics from your dynamic rule calculations. Lets say you did a historical data load and you don't want that to impact your averages. To do this, open the job in AWS Glue ETL and choose the **Data Quality** Tab, then choose **Statistics** and select the statistics that you want to exclude. You will be able to see a trend chart along with a table of statistics. Select the values you want to exclude and choose **Exclude Statistics**. Now the excluded statistics will not be included in the dynamic rule calculations.


Visual
Script
Job details
Runs
Data quality
Schedules
Version Control
Anomalies
0
Rules passed (0)
Rules failed (1)
0
Nov 08 21:38:51
Nov 08 23:50:01
Time (UTC)
Rules (1)
Statistics (1)
Anomalies (0)
All statistics
Filter by statistics
All statistics (1)
RowCount
default_context.RowCount
Statistic values and trend.
RowCount values
4
3.5
3
2.5
2
1.5
1
0.5
0
Nov 08 21:38:51
Nov 08 23:50:01
Time (UTC)
RowCount values
Prediction trend
Prediction upper bound
Prediction lower bound
RowCount (2)
Last updated (UTC) November 12, 2024 at 17:28:55
Save and retrain
Edit training inputs
Include statistics
Exclude statistics
Filter by statistic value
Filter by a date and time range

	Statistic value	Run date (UTC)	Training input
☑	4	November 8, 2024 at 23:50:01	Excluded statistic
☐	4	November 8, 2024 at 21:38:51	Included statistic

Analyzers

Note

Analyzers are not supported in AWS Glue Data Catalog.

DQDL rules use functions called *analyzers* to gather information about your data. This information is employed by a rule's Boolean expression to determine whether the rule should succeed or fail. For example, the RowCount rule `RowCount > 5` will use a row count analyzer to discover the number of rows in your dataset, and compare that count with the expression `> 5` to check whether more than five rows exist in the current dataset.

Sometimes, instead of authoring rules, we recommend creating analyzers and then have them generate statistics that can be used to detect anomalies. For such instances, you can create analyzers. Analyzers differ from rules in the following ways.

Characteristic	Analyzers	Rules
Part of ruleset	Yes	Yes
Generates statistics	Yes	Yes
Generates observations	Yes	Yes
Can evaluate and assert a condition	No	Yes
You can configure actions such as stop the jobs on failure, continue processing job	No	Yes

Analyzers can independently exist without rules, so you can quickly configure them and progressively build data quality rules.

Some rule types can be input in the `Analyzers` block of your ruleset to run the rules required for analyzers and gather information without applying checks for any condition. Some analyzers aren't associated with rules and can only be input in the `Analyzers` block. The following table indicates whether each item is supported as a rule or a standalone analyzer, along with additional details for each rule type.

Example Ruleset with Analyzer

The following ruleset uses:

- a dynamic rule to check whether a dataset is growing over its trailing average for the last three job runs
- a `DistinctValuesCount` analyzer to record the number of distinct values in the dataset's `Name` column
- a `ColumnLength` analyzer to track minimum and maximum `Name` size over time

Analyzer metric results can be viewed in the Data Quality tab for your job run.

```
Rules = [
   RowCount > avg(last(3))
]
Analyzers = [
   DistinctValuesCount "Name",
   ColumnLength "Name"
]
```

AWS Glue Data Quality supports the following analyzers.

Analyzer name	Functionality
`RowCount`	Computes row counts for a dataset
`Completeness`	Computes the completeness percentage of a column
`Uniqueness`	Computes the uniqueness percentage of a column
`Mean`	Computes mean of a numeric column
`Sum`	Computes sum of a numeric column
`StandardDeviation`	Computes standard deviation of a numeric column
`Entropy`	Computes entropy of a numeric column
`DistinctValuesCount`	Computes the number of distinct values in a column
`UniqueValueRatio`	Computes the unique values ratio in a column
`ColumnCount`	Computes the number of columns in a dataset
`ColumnLength`	Computes the length of a column

Analyzer name	Functionality
`ColumnValues`	Computes the minimum, maximum for numerical columns. Computes Minimum ColumnLength and Maximum ColumnLength for non-numeric columns
`ColumnCorrelation`	Computes column correlations for given columns
`CustomSql`	Computes statistics returned by the CustomSQL
`AllStatistics`	Computes the following statistics: • RowCount, ColumnCount • Every column: Completeness, uniqueness • Numerical: Min, Max, Entropy, Mean, Standard Dev, Sum • String: MinLength, MaxLength

Comments

You can use the '#' character to add a comment to your DQDL document. Anything after the '#' character and until the end of the line is ignored by DQDL.

```
Rules = [
    # More items should generally mean a higher price, so correlation should be
 positive
    ColumnCorrelation "price" "num_items" > 0
]
```

DQDL rule type reference

This section provides a reference for each rule type that AWS Glue Data Quality supports.

Note

- DQDL doesn't currently support nested or list-type column data.
- Bracketed values in the below table will be replaced with the information provided in rule arguments.
- Rules typically require an additional argument for expression.

Ruletyp	Descript on	Argume	Reporte Metrics	Support as Rule?	Support as Analyze	Returns row-level Results?	Dynamic rule support	Generat Observa ons	Supports Where Clause Syntax?
Aggrega Match	Checks if two datasets match by compari summar metrics like total sales amount. Useful for financial instituti ons to compare if all data is ingested from	One or more aggregat ons	When first and second aggregat on column names match: `Column. [C olumn]. gregate tch` When first and second aggregat on column	Yes	No	No	No	No	No

Ruletype	Description	Argument	Reported Metrics	Supported as Rule?	Supported as Analyzer?	Returns row-level Results?	Dynamic rule support?	Generates Observations	Supports Where Clause Syntax?
	source systems.		names different : `Column.[Column1,Column2].AggregateMatch`						

Ruletyp	Descript on	Argumen	Reporte Metrics	Support as Rule?	Support as Analyze	Returns row-level Results?	Dynami rule support	Generat Observa ons	Supports Where Clause Syntax?

AllStatistics	Standalo e analyzer to gather multiple metrics for the provided column in a dataset.	A single column name	For columns of all types:
			`Datase .RowCou`
			`Column. [C olumn]. mpleten s`
			`Column. [C olumn]. iquenes`
			Addition l metrics for string-valued columns
			`Columnl gth metrics`
				No	Yes	No	No	No	No

Ruletyp	Descript on	Argumen	Reporte Metrics	Support as Rule?	Support as Analyzer	Returns row-level Results?	Dynamic rule support	Generat Observa ons	Supports Where Clause Syntax?
			Addition l metrics for numeric-valued columns `ColumnV ues metrics`						
ColumnC relation	Checks how well two columns are correlate d.	Exactly two column names	`Multico mn.[Colum n1,Colu 2].Colu Correla on`	Yes	Yes	No	Yes	No	Yes
ColumnC nt	Checks if any columns are dropped	None	`Datase .Column unt`	Yes	Yes	No	Yes	Yes	No

Ruletyp	Descript on	Argume	Reporte Metrics	Support as Rule?	Support as Analyze	Returns row-level Results?	Dynamic rule support	Generat Observa ons	Supports Where Clause Syntax?
ColumnD aType	Checks if a column is complian with a datatype	Exactly one column name	`Column. [C olumn]. lumnDat ype.Con iance`	Yes	No	No	Yes, in row-level threshol expressio n	No	Yes
ColumnE sts	Checks if columns exist in a dataset. This allows custome building self service data platform to ensure certain columns are made available .	Exactly one column name	N/A	Yes	No	No	No	No	No

Ruletyp	Descript on	Argumen	Reporte Metrics	Support as Rule?	Support as Analyze	Returns row-level Results?	Dynamic rule support	Generat Observa ons	Supports Where Clause Syntax?

ColumnL gth	Checks if length of data is consistent.	Exactly one column name	`Column. [Column]. ximumLe th`
			`Column. [Column]. nimumLe th`
			Addition l metric when row-level threshol provided
			`Column. [Column]. lumnVa s.Comp nce`
				Yes	Yes	Yes, when row-level threshol provided	No	Yes. Only generate observat ons by analyzin Minimun and Maximur length	Yes

Ruletyp	Descript on	Argume	Reporte Metrics	Support as Rule?	Support as Analyze	Returns row-level Results?	Dynami rule support	Generat Observa ons	Supports Where Clause Syntax?
ColumnN esMatch ttern	Checks if column names match defined patterns Useful for governan e teams to enforce column name consisten cy.	A regex for column names	`Dataset .Column mesPatt nMatchR io`	Yes	No	No	No	No	No

Ruletyp	Descript on	Argumen	Reporte Metrics	Support as Rule?	Support as Analyze	Returns row-level Results?	Dynamic rule support	Generat Observa ons	Supports Where Clause Syntax?

ColumnV ues	Checks if data is consiste t per defined values. This rule supports regular expressi ns.	Exactly one column name	`Column. [C olumn]. ximum`
			`Column. [C olumn]. nimum`
			Addition l metric when row-level threshol provided
			`Column. [C olumn]. lumnVa s.Comp nce`
				Yes	Yes	Yes, when row-level threshol provided	No	Yes. Only generate observat ons by analyzin Minimu and Maximu values	Yes

Ruletyp	Descript on	Argume	Reporte Metrics	Support as Rule?	Support as Analyze	Returns row-level Results?	Dynamic rule support	Generat Observa ons	Supports Where Clause Syntax?
Complet ess	Checks for any blank or NULLs in data.	Exactly one column name	`Column. [Column]. mpleter s`	Yes	Yes	Yes	Yes	Yes	Yes

CustomS	Custome can impleme almost any type of data quality checks in SQL.	A SQL statemer
		(Optiona) A row-level threshol
			`Datase .Custor L`
			Addition l metric when row-level threshol provided
			`Datase .Custor L.Compl nce`
				Yes	No	Yes, when row-level threshol provided	Yes	No	No

DataFres ness	Checks if data is fresh.	Exactly one column name	`Column. [Column]. taFresh ss.Comp ance`	Yes	No	Yes	No	No	Yes

Ruletyp	Descript on	Argumer	Reporte Metrics	Support as Rule?	Support as Analyzer	Returns row-level Results?	Dynamic rule support?	Generat Observa ons	Supports Where Clause Syntax?
DatasetM tch	Compare two datasets and identifie s if they are in synch.	Name of a referenc dataset A column mapping (Optiona) Columns to check for matches	`Dataset [ Referer Dataset ias].Da setMatc`	Yes	No	Yes	Yes	No	No
DistinctV aluesCou t	Checks for duplicate values.	Exactly one column name	`Column. [C olumn]. stinctV uesCour`	Yes	Yes	Yes	Yes	Yes	Yes
DetectAr malies	Checks for anomalie in another rule type's reported metrics.	A rule type	Metric(s) reported by the rule type argumer	Yes	No	No	No	No	No

Ruletyp	Descript on	Argume	Reporte Metrics	Support as Rule?	Support as Analyze	Returns row-level Results?	Dynami rule support	Generat Observa ons	Supports Where Clause Syntax?
Entropy	Checks for entropy of the data.	Exactly one column name	`Column. [C olumn]. tropy`	Yes	Yes	No	Yes	No	Yes
IsComple e	Checks if 100% of the data is complet	Exactly one column name	`Column. [C olumn]. mpleter s`	Yes	No	Yes	No	No	Yes

IsPrimar Key	Checks if a column is a primary key (not NULL and unique).	Exactly one column name	For single column:
			`Column. [C olumn]. iquenes`
			For multiple columns
			`Multic mn. [Comma Delimi Columns Uniquer s`
				Yes	No	Yes	No	No	Yes

Ruletyp	Descript on	Argume	Reporte Metrics	Support as Rule?	Support as Analyze	Returns row-level Results?	Dynami rule support	Generat Observa ons	Supports Where Clause Syntax?
IsUnique	Checks if 100% of the data is unique.	Exactly one column name	`Column. [C olumn]. iquenes`	Yes	No	Yes	No	No	Yes
Mean	Checks if the mean matches the set threshol .	Exactly one column name	`Column. [C olumn]. an`	Yes	Yes	Yes	Yes	No	Yes
Referent alIntegri ty	Checks if two datasets have referential integrity .	One or more column names from dataset One or more column names from referenc dataset	`Column. [R eferenc atasetA as].Ref entialI egrity`	Yes	No	Yes	Yes	No	No

Ruletyp	Descript on	Argume	Reporte Metrics	Support as Rule?	Support as Analyze	Returns row-level Results?	Dynami rule support	Generat Observa ons	Supports Where Clause Syntax?
RowCou	Checks if record counts match a threshol .	None	`Dataset .RowCou`	Yes	Yes	No	Yes	Yes	Yes
RowCou atch	Checks if record counts between two datasets match.	Referenc dataset alias	`Dataset [ Referer Dataset ias].Rc ountMat`	Yes	No	No	Yes	No	No
Standar eviation	Checks if standard deviatio matches the threshol .	Exactly one column name	`Column [C olumn] andard iation`	Yes	Yes	Yes	Yes	No	Yes

Ruletyp	Descript on	Argumen	Reporte Metrics	Support as Rule?	Support as Analyze	Returns row-level Results?	Dynamic rule support	Generat Observa ons	Supports Where Clause Syntax?
SchemaM ch	Checks if schema between two datasets match.	Referenc dataset alias	`Dataset [ Referen Dataset ias].Sc maMatch`	Yes	No	No	Yes	No	No
Sum	Checks if sum matches a set threshol .	Exactly one column name	`Column. [C olumn]. m`	Yes	Yes	No	Yes	No	Yes
Uniquen s	Checks if uniquen s of dataset matches threshol .	Exactly one column name	`Column. [C olumn]. iquenes`	Yes	Yes	Yes	Yes	No	Yes
UniqueV ueRatio	Checks if the unique value ration matches threshol .	Exactly one column name	`Column. [C olumn]. iqueVal Ratio`	Yes	Yes	Yes	Yes	No	Yes

Ruletyp	Descript on	Argumen	Reporte Metrics	Support as Rule?	Support as Analyzer	Returns row-level Results?	Dynamic rule support	Generat Observa ons	Supports Where Clause Syntax?
FileFreshness	Checks if files in Amazon S3 are fresh.	File or Folder path and a threshol .	`Dataset .FileFr hness.C pliance` `Dataset .FileCo t`	Yes	No	No	No	No	No
FileMatc	Checks if contents of file match to a checksur or with other file. This rule uses checksur to validate if two files are same.	Source File or Folder path and Target file or folder path.	No statistic s are generate .	Yes	No	No	No	No	No

Ruletyp	Descript on	Argumer	Reporte Metrics	Support as Rule?	Support as Analyzer	Returns row-level Results?	Dynamic rule support	Generat Observa ons	Supports Where Clause Syntax?

FileSize	Checks if the size of a file matches with a specified conditior .	File or folder path and threshol .	`Dataset .FileSi .Compli ce`
			`Dataset .FileC t`
			`Dataset .Maximu ileSize`
			`Dataset .Minimu ileSize`
				Yes	No	No	No	No	No

FileUniq eness	Checks if files are unique using checksur .	File or folder path and threshol .	`Dataset .FileU uenessR io` `Dataset .FileC t`	Yes	No	No	No	No	No

Topics

- AggregateMatch
- ColumnCorrelation
- ColumnCount
- ColumnDataType

- ColumnExists
- ColumnLength
- ColumnNamesMatchPattern
- ColumnValues
- Completeness
- CustomSQL
- DataFreshness
- DatasetMatch
- DistinctValuesCount
- Entropy
- IsComplete
- IsPrimaryKey
- IsUnique
- Mean
- ReferentialIntegrity
- RowCount
- RowCountMatch
- StandardDeviation
- Sum
- SchemaMatch
- Uniqueness
- UniqueValueRatio
- DetectAnomalies
- FileFreshness
- FileMatch
- FileUniqueness
- FileSize

AggregateMatch

Checks the ratio of two column aggregations against a given expression. This ruletype works on multiple datasets. The two column aggregations are evaluated and a ratio is produced by dividing the result of the first column aggregation with the result of the second column aggregation. The ratio is checked against the provided expression to produce a boolean response.

Syntax

Column aggregation

```
AggregateMatch <AGG_OPERATION> (<OPTIONAL_REFERENCE_ALIAS>.<COL_NAME>)
```

- **AGG_OPERATION** – The operation to use for the aggregation. Currently, `sum` and `avg` are supported.

  Supported column types: Byte, Decimal, Double, Float, Integer, Long, Short
- **OPTIONAL_REFERENCE_ALIAS** – This parameter needs to be provided if the column is from a reference dataset and not the primary dataset. If you are using this rule in the AWS Glue Data Catalog, your reference alias must follow the format "<database_name>.<table_name>.<column_name>

  Supported column types: Byte, Decimal, Double, Float, Integer, Long, Short
- **COL_NAME** – The name of the column to aggregate.

  Supported column types: Byte, Decimal, Double, Float, Integer, Long, Short

Example: Average

```
"avg(rating)"
```

Example: Sum

```
"sum(amount)"
```

Example: Average of column in reference dataset

```
"avg(reference.rating)"
```

Rule

```
AggregateMatch <AGG_EXP_1> <AGG_EXP_2> <EXPRESSION>
```

- **AGG_EXP_1** – The first column aggregation.

  Supported column types: Byte, Decimal, Double, Float, Integer, Long, Short

  Supported column types: Byte, Decimal, Double, Float, Integer, Long, Short
- **AGG_EXP_2** – The second column aggregation.

  Supported column types: Byte, Decimal, Double, Float, Integer, Long, Short

  Supported column types: Byte, Decimal, Double, Float, Integer, Long, Short
- **EXPRESSION** – An expression to run against the rule type response in order to produce a Boolean value. For more information, see Expressions.

Example: Aggregate Match using sum

The following example rule checks whether the sum of the values in the `amount` column is exactly equal to the sum of the values in the `total_amount` column.

```
AggregateMatch "sum(amount)" "sum(total_amount)" = 1.0
```

Example: Aggregate Match using average

The following example rule checks whether the average of the values in the `ratings` column is equal to at least 90% of the average of the values in the `ratings` column in the `reference` dataset. The reference dataset is provided as an additional data source in the ETL or Data Catalog experience.

In AWS Glue ETL, you can use:

```
AggregateMatch "avg(ratings)" "avg(reference.ratings)" >= 0.9
```

In the AWS Glue Data Catalog, you can use:

```
AggregateMatch "avg(ratings)" "avg(database_name.tablename.ratings)" >= 0.9
```

Null behavior

The `AggregateMatch` rule will ignore rows with NULL values in the calculation of the aggregation methods (sum/mean). For example:

```
+---+-----------+
|id |units      |
+---+-----------+
|100|0          |
|101|null       |
|102|20         |
|103|null       |
|104|40         |
+---+-----------+
```

The mean of column `units` will be (0 + 20 + 40) / 3 = 20. Rows 101 and 103 are not considered in this calculation.

ColumnCorrelation

Checks the *correlation* between two columns against a given expression. AWS Glue Data Quality uses the Pearson correlation coefficient to measure the linear correlation between two columns. The result is a number between -1 and 1 that measures the strength and direction of the relationship.

Syntax

```
ColumnCorrelation <COL_1_NAME> <COL_2_NAME> <EXPRESSION>
```

- **COL_1_NAME** – The name of the first column that you want to evaluate the data quality rule against.

  Supported column types: Byte, Decimal, Double, Float, Integer, Long, Short
- **COL_2_NAME** – The name of the second column that you want to evaluate the data quality rule against.

  Supported column types: Byte, Decimal, Double, Float, Integer, Long, Short
- **EXPRESSION** – An expression to run against the rule type response in order to produce a Boolean value. For more information, see Expressions.

Example: Column correlation

The following example rule checks whether the correlation coefficient between the columns `height` and `weight` has a strong positive correlation (a coefficient value greater than 0.8).

```
ColumnCorrelation "height" "weight" > 0.8
```

```
ColumnCorrelation "weightinkgs" "Salary" > 0.8 where "weightinkgs > 40"
```

Sample dynamic rules

- `ColumnCorrelation "colA" "colB" between min(last(10)) and max(last(10))`
- `ColumnCorrelation "colA" "colB" < avg(last(5)) + std(last(5))`

Null behavior

The `ColumnCorrelation` rule will ignore rows with `NULL` values in the calculation of the correlation. For example:

```
+---+-----------+
|id |units      |
+---+-----------+
|100|0          |
|101|null       |
|102|20         |
|103|null       |
|104|40         |
+---+-----------+
```

Rows 101 and 103 will be ignored, and the `ColumnCorrelation` will be 1.0.

ColumnCount

Checks the column count of the primary dataset against a given expression. In the expression, you can specify the number of columns or a range of columns using operators like > and <.

Syntax

```
ColumnCount <EXPRESSION>
```

- **EXPRESSION** – An expression to run against the rule type response in order to produce a Boolean value. For more information, see Expressions.

Example: Column count numeric check

The following example rule checks whether the column count is within a given range.

```
ColumnCount between 10 and 20
```

Sample dynamic rules

- `ColumnCount >= avg(last(10))`
- `ColumnCount between min(last(10))-1 and max(last(10))+1`

ColumnDataType

Checks if the values in a given column can be cast in Apache Spark to the provided type. Accepts a `with threshold` expression to check for a subset of the values in the column.

Syntax

```
ColumnDataType <COL_NAME> = <EXPECTED_TYPE>
```

- **COL_NAME** – The name of the column that you want to evaluate the data quality rule against.

  Supported column types: String type

  Supported column types: Byte, Decimal, Double, Float, Integer, Long, Short
- **EXPECTED_TYPE** – The expected type of the values in the column.

  Supported values: Boolean, Date, Timestamp, Integer, Double, Float, Long

  Supported column types: Byte, Decimal, Double, Float, Integer, Long, Short
- **EXPRESSION** – An optional expression to specify the percentage of values that should be of the expected type.

  Supported column types: Byte, Decimal, Double, Float, Integer, Long, Short

Example: Column data type integers as strings

The following example rule checks whether the values in the given column, which is of type string, can be cast as integers.

```
ColumnDataType "colA" = "INTEGER"
```

Example: Column data type integers as strings check for a subset of the values

The following example rule checks whether more than 90% of the values in the given column, which is of type string, can be cast as integers.

```
ColumnDataType "colA" = "INTEGER" with threshold > 0.9
```

ColumnExists

Checks whether a column exists.

Syntax

```
ColumnExists <COL_NAME>
```

- **COL_NAME** – The name of the column that you want to evaluate the data quality rule against.

  Supported column types: Any column type

Example: Column exists

The following example rule checks whether the column named `Middle_Name` exists.

```
ColumnExists "Middle_Name"
```

ColumnLength

Checks whether the length of each row in a column conforms to a given expression.

Syntax

```
ColumnLength <COL_NAME><EXPRESSION>
```

- **COL_NAME** – The name of the column that you want to evaluate the data quality rule against.

  Supported column types: String
- **EXPRESSION** – An expression to run against the rule type response in order to produce a Boolean value. For more information, see Expressions.

Example: Column row length

The following example rule checks whether the value in each row in the column named `Postal_Code` is 5 characters long.

```
ColumnLength "Postal_Code" = 5
ColumnLength "weightinkgs" = 2 where "weightinkgs > 10"
```

Null behavior

The `ColumnLength` rule treats `NULL`s as 0 length strings. For a `NULL` row:

```
ColumnLength "Postal_Code" > 4 # this will fail
```

```
ColumnLength "Postal_Code" < 6 # this will succeed
```

The following example compound rule provides a way to explicitly fail `NULL` values:

```
(ColumnLength "Postal_Code" > 4) AND (ColumnValues "Postal_Code" != NULL)
```

ColumnNamesMatchPattern

Checks whether the names of all columns in the primary dataset match the given regular expression.

Syntax

```
ColumnNamesMatchPattern <PATTERN>
```

- **PATTERN** – The pattern you want to evaluate the data quality rule against.

  Supported column types: Byte, Decimal, Double, Float, Integer, Long, Short

Example: Column names match pattern

The following example rule checks whether all columns start with the prefix "aws_"

```
ColumnNamesMatchPattern "aws_.*"
ColumnNamesMatchPattern "aws_.*" where "weightinkgs > 10"
```

ColumnValues

Runs an expression against the values in a column.

Syntax

```
ColumnValues <COL_NAME> <EXPRESSION>
```

- **COL_NAME** – The name of the column that you want to evaluate the data quality rule against.

  Supported column types: Any column type
- **EXPRESSION** – An expression to run against the rule type response in order to produce a Boolean value. For more information, see Expressions.

Example: Allowed values

The following example rule checks whether each value in the specified column is in a set of allowed values (including null, empty, and strings with only whitespaces).

```
ColumnValues "Country" in [ "US", "CA", "UK", NULL, EMPTY, WHITESPACES_ONLY ]
ColumnValues "gender" in ["F", "M"] where "weightinkgs < 10"
```

Example: Regular expression

The following example rule checks the values in a column against a regular expression.

```
ColumnValues "First_Name" matches "[a-zA-Z]*"
```

Example: Date values

The following example rule checks the values in a date column against a date expression.

```
ColumnValues "Load_Date" > (now() - 3 days)
```

Example: Numeric values

The following example rule checks whether the column values match a certain numeric constraint.

```
ColumnValues "Customer_ID" between 1 and 2000
```

Null behavior

For all `ColumnValues` rules (other than `!=` and `NOT IN`), `NULL` rows will fail the rule. If the rule fails due to a null value, the failure reason will display the following:

```
Value: NULL does not meet the constraint requirement!
```

The following example compound rule provides a way to explicitly allow for `NULL` values:

```
(ColumnValues "Age" > 21) OR (ColumnValues "Age" = NULL)
```

Negated ColumnValues rules using the `!=` and `not in` syntax will pass for `NULL` rows. For example:

```
ColumnValues "Age" != 21
```

```
ColumnValues "Age" not in [21, 22, 23]
```

The following examples provide a way to explicitly fail `NULL` values

```
(ColumnValues "Age" != 21) AND (ColumnValues "Age" != NULL)
```

```
ColumnValues "Age" not in [21, 22, 23, NULL]
```

Completeness

Checks the percentage of complete (non-null) values in a column against a given expression.

Syntax

```
Completeness <COL_NAME> <EXPRESSION>
```

- **COL_NAME** – The name of the column that you want to evaluate the data quality rule against.

  Supported column types: Any column type
- **EXPRESSION** – An expression to run against the rule type response in order to produce a Boolean value. For more information, see Expressions.

Example: Null value percentage

The following example rules check if more than 95 percent of the values in a column are complete.

```
Completeness "First_Name" > 0.95
Completeness "First_Name" > 0.95 where "weightinkgs > 10"
```

Sample dynamic rules

- `Completeness "colA" between min(last(5)) - 1 and max(last(5)) + 1`
- `Completeness "colA" <= avg(last(10))`

Null behavior

Note on CSV Data Formats: Blank rows on CSV columns can display multiple behaviors.

- If a column is of `String` type, the blank row will be recognized as an empty string and will not fail the `Completeness` rule.
- If a column is of another data type like `Int`, the blank row will be recognized as `NULL` and will fail the `Completeness` rule.

CustomSQL

This rule type has been extended to support two use cases:

- Run a custom SQL statement against a dataset and checks the return value against a given expression.

- Run a custom SQL statement where you specify a column name in your SELECT statement against which you compare with some condition to get row-level results.

Syntax

```
CustomSql <SQL_STATEMENT> <EXPRESSION>
```

- **SQL_STATEMENT** – A SQL statement that returns a single numeric value, surrounded by double quotes.
- **EXPRESSION** – An expression to run against the rule type response in order to produce a Boolean value. For more information, see Expressions.

Example: Custom SQL to retrieve an overall rule outcome

This example rule uses a SQL statement to retrieve the record count for a data set. The rule then checks that the record count is between 10 and 20.

```
CustomSql "select count(*) from primary" between 10 and 20
```

Example: Custom SQL to retrieve row-level results

This example rule uses a SQL statement wherein you specify a column name in your SELECT statement against which you compare with some condition to get row level results. A threshold condition expression defines a threshold of how many records should fail for the entire rule to fail. Note that a rule may not contain both a condition and keyword together.

```
CustomSql "select Name from primary where Age  > 18"
```

or

```
CustomSql "select Name from primary where Age > 18" with threshold  > 3
```

Important

The `primary` alias stands in for the name of the data set that you want to evaluate. When you work with visual ETL jobs on the console, `primary` always represents the `DynamicFrame` being passed to the `EvaluateDataQuality.apply()` transform. When you use the AWS Glue Data Catalog to run data quality tasks against a table, `primary` represents the table.

If you are in AWS Glue Data Catalog, you can also use the actual table names:

```
CustomSql "select count(*) from database.table" between 10 and 20
```

You can also join multiple tables to compare different data elements:

```
CustomSql "select count(*) from database.table inner join database.table2 on id1 = id2"
 between 10 and 20
```

In AWS Glue ETL, CustomSQL can identify records that failed the data quality checks. For this to work, you need to return records that are part of the primary table for which you are evaluating data quality. Records that are returned as part of the query are considered successful and records that are not returned are considered failed. This works by joining the result of your CustomSQL query with the original dataset. There may be performance implications based on the complexity of your SQL query.

To do this:

- You need to select at least 1 column from your primary table.
  - `select count(*) from primary` is a valid query for OVERALL CustomSQL DQ rule but not for Row Level Custom SQL.
  - This rule will throw an error during evaluation: `The output from CustomSQL must contain at least one column that matches the input dataset for AWS Glue Data Quality to provide row level results. The SQL query is a valid query but the columns from the SQL result are not present in the Input Dataset. Ensure that matching columns are returned from the SQL.`
- In your SQL query, select a `Primary Key` from your table or select a set of columns that form a composite key. Not doing so may result in inconsistent results due to matching of duplicate rows and degraded performance.
- Select keys ONLY from your primary table and not from your reference tables.

The following rule will ensure that records with age < 100 are identified as successful and records that are above are marked as failed.

```
CustomSql "select id from primary where age < 100"
```

This CustomSQL rule will pass when 50% of the records have age > 10 and will also identify records that failed. The records returned by this CustomSQL will be considered passed while the ones not returned will be considered failed.

```
CustomSQL "select ID, CustomerID from primary where age > 10" with threshold > 0.5
```

Note: CustomSQL rule will fail if you return records that are not available in the dataset.

DataFreshness

Checks the freshness of data in a column by evaluating the difference between the current time and the values of a date column. You can specify a time-based expression for this rule type to make sure that column values are up to date.

Syntax

```
DataFreshness <COL_NAME> <EXPRESSION>
```

- **COL_NAME** – The name of the column that you want to evaluate the data quality rule against.

  Supported column types: Date
- **EXPRESSION** – A numeric expression in hours or days. You must specify the time unit in your expression.

Example: Data freshness

The following example rules check for data freshness.

```
DataFreshness "Order_Date" <= 24 hours
DataFreshness "Order_Date" between 2 days and 5 days
```

Null behavior

The `DataFreshness` rules will fail for rows with `NULL` values. If the rule fails due to a null value, the failure reason will display the following:

```
80.00 % of rows passed the threshold
```

where 20% of the rows that failed include the rows with NULL.

The following example compound rule provides a way to explicitly allow for NULL values:

```
(DataFreshness "Order_Date" <= 24 hours) OR (ColumnValues "Order_Date" = NULL)
```

Data Freshness for Amazon S3 objects

Sometimes you will need to validate the freshness of data based on the Amazon S3 file creating time. To do this, you can use the following code to get the timestamp and add it to your dataframe, and then apply Data Freshness checks.

```
df = glueContext.create_data_frame.from_catalog(database = "default", table_name =
 "mytable")
df = df.withColumn("file_ts", df["_metadata.file_modification_time"])

Rules = [
 DataFreshness "file_ts" < 24 hours
]
```

DatasetMatch

Checks if the data in the primary dataset matches the data in a reference dataset. The two datasets are joined using the provided key column mappings. Additional column mappings can be provided should you wish to check for the equality of the data in only those columns. Note that for **DataSetMatch** to work, your join keys should be unique and should not be NULL (must be a primary key). If you don’t satisfy these conditions, you will get the error message, “Provided key map not suitable for given data frames”. In cases where you can’t have joined keys that are unique, consider using other ruletypes such as **AggregateMatch** to match on summary data.

Syntax

```
DatasetMatch <REFERENCE_DATASET_ALIAS> <JOIN CONDITION WITH
 MAPPING> <OPTIONAL_MATCH_COLUMN_MAPPINGS> <EXPRESSION>
```

- **REFERENCE_DATASET_ALIAS** – The alias of the reference dataset with which you compare data from the primary dataset.
- **KEY_COLUMN_MAPPINGS** – A comma-separated list of column names that form a key in the datasets. If the column names are not the same in both datasets, you must separated them with a `->`

- **OPTIONAL_MATCH_COLUMN_MAPPINGS** – You can supply this parameter if you want to check for matching data only in certain columns. It uses the same syntax as the key column mappings. If this parameter is not provided, we will match the data in all remaining columns. The remaining, non-key columns must have the same names in both datasets.
- **EXPRESSION** – An expression to run against the rule type response in order to produce a Boolean value. For more information, see Expressions.

Example: Match set datasets using ID column

The following example rule checks that more than 90% of the primary dataset matches the reference dataset, using the "ID" column to join the two datasets. It compares all columns in this case.

```
DatasetMatch "reference" "ID" >= 0.9
```

Example: Match set datasets using multiple key columns

In the following example, the primary dataset and the reference dataset have different names for the key columns. `ID_1` and `ID_2` together form a composite key in the primary dataset. `ID_ref1` and `ID_ref2` together forms a composite key in the reference dataset. In this scenario, you can use the special syntax to supply the column names.

```
DatasetMatch "reference" "ID_1->ID_ref1,ID_ref2->ID_ref2" >= 0.9
```

Example: Match set datasets using multiple key columns and check that specific column matches

This example builds on the previous example. We want to check that only the column containing the amounts match. This column is named `Amount1` in the primary dataset and `Amount2` in the reference dataset. You want an exact match.

```
DatasetMatch "reference" "ID_1->ID_ref1,ID_2->ID_ref2" "Amount1->Amount2" >= 0.9
```

DistinctValuesCount

Checks the number of distinct values in a column against a given expression.

Syntax

```
DistinctValuesCount <COL_NAME> <EXPRESSION>
```

- **COL_NAME** – The name of the column that you want to evaluate the data quality rule against.

  Supported column types: Any column type
- **EXPRESSION** – An expression to run against the rule type response in order to produce a Boolean value. For more information, see Expressions.

Example: Distinct column value count

The following example rule checks that the column named `State` contains more than 3 distinct values.

```
DistinctValuesCount "State" > 3
DistinctValuesCount "Customer_ID" < 6  where "Customer_ID < 10"
```

Sample dynamic rules

- `DistinctValuesCount "colA" between avg(last(10))-1 and avg(last(10))+1`
- `DistinctValuesCount "colA" <= index(last(10),2) + std(last(5))`

Entropy

Checks whether the *entropy* value of a column matches a given expression. Entropy measures the level of information that's contained in a message. Given the probability distribution over values in a column, entropy describes how many bits are required to identify a value.

Syntax

```
Entropy <COL_NAME> <EXPRESSION>
```

- **COL_NAME** – The name of the column that you want to evaluate the data quality rule against.

  Supported column types: Any column type
- **EXPRESSION** – An expression to run against the rule type response in order to produce a Boolean value. For more information, see Expressions.

Example: Column entropy

The following example rule checks that the column named `Feedback` has an entropy value greater than one.

```
Entropy "Star_Rating" > 1
Entropy "First_Name" > 1  where "Customer_ID < 10"
```

Sample dynamic rules

- `Entropy "colA" < max(last(10))`
- `Entropy "colA" between min(last(10)) and max(last(10))`

IsComplete

Checks whether all of the values in a column are complete (non-null).

Syntax

```
IsComplete <COL_NAME>
```

- **COL_NAME** – The name of the column that you want to evaluate the data quality rule against.

  Supported column types: Any column type

Example: Null values

The following example checks whether all of the values in a column named `email` are non-null.

```
IsComplete "email"
IsComplete "Email" where "Customer_ID between 1 and 50"
IsComplete "Customer_ID"  where "Customer_ID < 16 and Customer_ID != 12"
IsComplete "passenger_count" where "payment_type<>0"
```

Null behavior

Note on CSV Data Formats: Blank rows on CSV columns can display multiple behaviors.

- If a column is of `String` type, the blank row will be recognized as an empty string and will not fail the `Completeness` rule.
- If a column is of another data type like `Int`, the blank row will be recognized as `NULL` and will fail the `Completeness` rule.

IsPrimaryKey

Checks whether a column contains a primary key. A column contains a primary key if all of the values in the column are unique and complete (non-null). You can also check for primary keys with multiple columns.

Syntax

```
IsPrimaryKey <COL_NAME>
```

- **COL_NAME** – The name of the column that you want to evaluate the data quality rule against.

  Supported column types: Any column type

Example: Primary key

The following example rule checks whether the column named `Customer_ID` contains a primary key.

```
IsPrimaryKey "Customer_ID"
IsPrimaryKey "Customer_ID" where "Customer_ID < 10"
```

Example: Primary key with multiple columns. Any of the following examples are valid.

```
IsPrimaryKey "colA" "colB"
IsPrimaryKey "colA" "colB" "colC"
IsPrimaryKey colA "colB" "colC"
```

IsUnique

Checks whether all of the values in a column are unique, and returns a Boolean value.

Syntax

```
IsUnique <COL_NAME>
```

- **COL_NAME** – The name of the column that you want to evaluate the data quality rule against.

  Supported column types: Any column type

Examples

The following example rule checks whether all of the values in a column named `email` are unique.

```
IsUnique "email"
IsUnique "Customer_ID" where "Customer_ID < 10"]
```

The following example rule checks multiple columns.

```
IsUnique "vendorid" "tpep_pickup_datetime"
```

Mean

Checks whether the mean (average) of all the values in a column matches a given expression.

Syntax

```
Mean <COL_NAME> <EXPRESSION>
```

- **COL_NAME** – The name of the column that you want to evaluate the data quality rule against.

  Supported column types: Byte, Decimal, Double, Float, Integer, Long, Short
- **EXPRESSION** – An expression to run against the rule type response in order to produce a Boolean value. For more information, see Expressions.

Example: Average value

The following example rule checks whether the average of all of the values in a column exceeds a threshold.

```
Mean "Star_Rating" > 3
Mean "Salary" < 6200 where "Customer_ID < 10"
```

Sample dynamic rules

- `Mean "colA" > avg(last(10)) + std(last(2))`
- `Mean "colA" between min(last(5)) - 1 and max(last(5)) + 1`

Null behavior

The `Mean` rule will ignore rows with `NULL` values in the calculation of the mean. For example:

```
+---+-----------+
|id |units      |
+---+-----------+
|100|0          |
|101|null       |
|102|20         |
|103|null       |
|104|40         |
+---+-----------+
```

The mean of column `units` will be (0 + 20 + 40) / 3 = 20. Rows 101 and 103 are not considered in this calculation.

ReferentialIntegrity

Checks to what extent the values of a set of columns in the primary dataset are a subset of the values of a set of columns in a reference dataset.

Syntax

```
ReferentialIntegrity <PRIMARY_COLS> <REFERENCE_DATASET_COLS> <EXPRESSION>
```

- **PRIMARY_COLS** – A comma-separated list of column names in the primary dataset.

  Supported column types: Byte, Decimal, Double, Float, Integer, Long, Short
- **REFERENCE_DATASET_COLS** – This parameter contains two parts separated by a period. The first part is the alias of the reference dataset. The second part is the comma-separated list of column names in the reference dataset enclosed in braces.

  Supported column types: Byte, Decimal, Double, Float, Integer, Long, Short

- **EXPRESSION** – An expression to run against the rule type response in order to produce a Boolean value. For more information, see Expressions.

Example: Check the referential integrity of a zip code column

The following example rule checks that more than 90% of the values in the `zipcode` column in the primary dataset, are present in the `zipcode` column in the `reference` dataset.

```
ReferentialIntegrity "zipcode" "reference.zipcode" >= 0.9
```

Example: Check the referential integrity of the city and state columns

In the following example, columns containing city and state information exist in the primary dataset and the reference dataset. The names of the columns are different in both datasets. The rule checks if the set of values of the columns in the primary dataset is exactly equal to the set of values of the columns in the reference dataset.

```
ReferentialIntegrity "city,state" "reference.{ref_city,ref_state}" = 1.0
```

Sample dynamic rules

- `ReferentialIntegrity "city,state" "reference.{ref_city,ref_state}" > avg(last(10))`
- `ReferentialIntegrity "city,state" "reference.{ref_city,ref_state}" between min(last(10)) - 1 and max(last(10)) + 1`

RowCount

Checks the row count of a dataset against a given expression. In the expression, you can specify the number of rows or a range of rows using operators like > and <.

Syntax

```
RowCount <EXPRESSION>
```

- **EXPRESSION** – An expression to run against the rule type response in order to produce a Boolean value. For more information, see Expressions.

Example: Row count numeric check

The following example rule checks whether the row count is within a given range.

```
RowCount between 10 and 100
RowCount between 1 and 50 where "Customer_ID < 10"
```

Sample dynamic rules

```
RowCount > avg(last(10)) *0.8
```

RowCountMatch

Checks the ratio of the row count of the primary dataset and the row count of a reference dataset against the given expression.

Syntax

```
RowCountMatch <REFERENCE_DATASET_ALIAS> <EXPRESSION>
```

- **REFERENCE_DATASET_ALIAS –** The alias of the reference dataset against which to compare row counts.

  Supported column types: Byte, Decimal, Double, Float, Integer, Long, Short
- **EXPRESSION –** An expression to run against the rule type response in order to produce a Boolean value. For more information, see Expressions.

Example: Row count check against a reference dataset

The following example rule checks whether the row count of the primary dataset is at least 90% of the row count of the reference dataset.

```
RowCountMatch "reference" >= 0.9
```

StandardDeviation

Checks the standard deviation of all of the values in a column against a given expression.

Syntax

```
StandardDeviation <COL_NAME> <EXPRESSION>
```

- **COL_NAME** – The name of the column that you want to evaluate the data quality rule against.

  Supported column types: Byte, Decimal, Double, Float, Integer, Long, Short
- **EXPRESSION** – An expression to run against the rule type response in order to produce a Boolean value. For more information, see Expressions.

Example: Standard deviation

The following example rule checks whether the standard deviation of the values in a column named `colA` is less than a specified value.

```
StandardDeviation "Star_Rating" < 1.5
StandardDeviation "Salary" < 3500 where "Customer_ID < 10"
```

Sample dynamic rules

- `StandardDeviation "colA" > avg(last(10) + 0.1`
- `StandardDeviation "colA" between min(last(10)) - 1 and max(last(10)) + 1`

Null behavior

The `StandardDeviation` rule will ignore rows with `NULL` values in the calculation of standard deviation. For example:

```
+---+-----------+-----------+
|id |units1     |units2     |
+---+-----------+-----------+
|100|0          |0          |
|101|null       |0          |
|102|20         |20         |
|103|null       |0          |
|104|40         |40         |
+---+-----------+-----------+
```

The standard deviation of column `units1` will not consider rows 101 and 103 and result to 16.33. The standard deviation for column `units2` will result in 16.

Sum

Checks the sum of all the values in a column against a given expression.

Syntax

```
Sum <COL_NAME> <EXPRESSION>
```

- **COL_NAME** – The name of the column that you want to evaluate the data quality rule against.

  Supported column types: Byte, Decimal, Double, Float, Integer, Long, Short
- **EXPRESSION** – An expression to run against the rule type response in order to produce a Boolean value. For more information, see Expressions.

Example: Sum

The following example rule checks whether the sum of all of the values in a column exceeds a given threshold.

```
Sum "transaction_total" > 500000
Sum "Salary" < 55600 where "Customer_ID < 10"
```

Sample dynamic rules

- `Sum "ColA" > avg(last(10))`
- `Sum "colA" between min(last(10)) - 1 and max(last(10)) + 1`

Null behavior

The `Sum` rule will ignore rows with `NULL` values in the calculation of sum. For example:

```
+---+-----------+
|id |units      |
+---+-----------+
|100|0          |
|101|null       |
|102|20         |
|103|null       |
|104|40         |
+---+-----------+
```

The sum of column `units` will not consider rows 101 and 103 and result to (0 + 20 + 40) = 60.

SchemaMatch

Checks if the schema of the primary dataset matches the schema of a reference dataset. The schema check is done column by column. The schema of two columns match if the names are identical and the types are identical. The order of the columns does not matter.

Syntax

```
SchemaMatch <REFERENCE_DATASET_ALIAS> <EXPRESSION>
```

- **REFERENCE_DATASET_ALIAS** – The alias of the reference dataset against which to compare schemas.

  Supported column types: Byte, Decimal, Double, Float, Integer, Long, Short
- **EXPRESSION** – An expression to run against the rule type response in order to produce a Boolean value. For more information, see Expressions.

Example: SchemaMatch

The following example rule checks whether the schema of the primary dataset exactly matches the schema of a reference dataset.

```
SchemaMatch "reference" = 1.0
```

Uniqueness

Checks the percentage of unique values in a column against a given expression. Unique values occur exactly once.

Syntax

```
Uniqueness <COL_NAME> <EXPRESSION>
```

- **COL_NAME** – The name of the column that you want to evaluate the data quality rule against.

Supported column types: Any column type

- **EXPRESSION** – An expression to run against the rule type response in order to produce a Boolean value. For more information, see Expressions.

Example

The following example rule checks whether the percentage of unique values in a column matches certain numeric criteria.

```
Uniqueness "email" = 1.0
Uniqueness "Customer_ID" != 1.0 where "Customer_ID < 10"
```

The following example rule checks multiple columns.

```
Uniqueness "vendorid" "tpep_pickup_datetime" = 1
```

Sample dynamic rules

- `Uniqueness "colA" between min(last(10)) and max(last(10))`
- `Uniqueness "colA" >= avg(last(10))`

UniqueValueRatio

Checks the *unique value ratio* of a column against a given expression. A unique value ratio is the fraction of unique values divided by the number of all distinct values in a column. Unique values occur exactly one time, while distinct values occur *at least* once.

For example, the set `[a, a, b]` contains one unique value (b) and two distinct values (a and b). So the unique value ratio of the set is ½ = 0.5.

Syntax

```
UniqueValueRatio <COL_NAME> <EXPRESSION>
```

- **COL_NAME** – The name of the column that you want to evaluate the data quality rule against.

Supported column types: Any column type

- **EXPRESSION** – An expression to run against the rule type response in order to produce a Boolean value. For more information, see Expressions.

Example: Unique value ratio

This example checks the unique value ratio of a column against a range of values.

```
UniqueValueRatio "test_score" between 0 and 0.5
UniqueValueRatio "Customer_ID" between 0 and 0.9 where "Customer_ID < 10"
```

Sample dynamic rules

- `UniqueValueRatio "colA" > avg(last(10))`
- `UniqueValueRatio "colA" <= index(last(10),2) + std(last(5))`

DetectAnomalies

Detects anomalies for a given data quality rule. Every execution of DetectAnomalies rule result in saving evaluated value for the given rule. When there is enough data gathered, anomaly detection algorithm takes all historical data for that given rule and runs anomaly detection. DetectAnomalies rule fails when anomaly is detected. More info about what anomaly was detected can be obtained from Observations.

Syntax

```
        DetectAnomalies <RULE_NAME> <RULE_PARAMETERS>
```

RULE_NAME – The name of the rule that you want to evaluate and detect anomalies for. Supported rules:

- "RowCount"
- "Completeness"
- "Uniqueness"

- "Mean"
- "Sum"
- "StandardDeviation"
- "Entropy"
- "DistinctValuesCount"
- "UniqueValueRatio"
- "ColumnLength"
- "ColumnValues"
- "ColumnCorrelation"
- "CustomSQL"
- "ColumnCount"

`RULE_PARAMETERS` – some rules require additional parameters to run. Refer to the given rule documentation to see required parameters.

Example: Anomalies for RowCount

For example, if we want to detect RowCount anomalies, we provide RowCount as a rule name.

```
DetectAnomalies "RowCount"
```

Example: Anomalies for ColumnLength

For example, if we want to detect ColumnLength anomalies, we provide ColumnLength as a rule name and the column name.

```
DetectAnomalies "ColumnLength" "id"
```

FileFreshness

FileFreshness ensures your data files are fresh based on the condition you provide. It uses your files' last modified time to ensure that data files or the entire folder is up-to-date.

This rule gathers two metrics:

- FileFreshness compliance based on the rule you set up

- The number of files that were scanned by the rule

```
{"Dataset.*.FileFreshness.Compliance":1,"Dataset.*.FileCount":1}
```

Anomaly detection does not consider these metrics.

Checking file freshness

The following rule ensures that tickets.parquet was created in the past 24 hours.

```
FileFreshness "s3://amzn-s3-demo-bucket/artifacts/file/tickets/tickets.parquet" >
 (now() - 24 hours)
```

Checking folder freshness

The following rule passes if all files in the folder were created or modified in past 24 hours.

```
FileFreshness "s3://bucket/" >= (now() -1 days)
FileFreshness "s3://amzn-s3-demo-bucket/artifacts/file/tickets/" >= (now() - 24 hours)
```

Checking folder or file freshness with threshold

The following rule passes if 10% of the files in the folder “tickets“ were created or modified in the past 10 days.

```
FileFreshness "s3://amzn-s3-demo-bucket/artifacts/file/tickets/" < (now() - 10 days)
 with threshold > 0.1
```

Checking files or folders with specific dates

You can check for file freshness for specific days.

```
FileFreshness "s3://amzn-s3-demo-bucket/artifacts/file/tickets/" > "2020-01-01"
FileFreshness "s3://amzn-s3-demo-bucket/artifacts/file/tickets/" between "2023-01-01"
 and "2024-01-01"
```

Checking files or folders with time

You can use FileFreshness to ensure that files have arrived based on certain times.

```
FileFreshness "s3://amzn-s3-demo-bucket/artifacts/file/tickets/" between now() and
 (now() - 45 minutes)
FileFreshness "s3://amzn-s3-demo-bucket/artifacts/file/tickets/" between "9:30 AM" and
 "9:30 PM"
FileFreshness "s3://amzn-s3-demo-bucket/artifacts/file/tickets/" > (now() - 10 minutes)
FileFreshness "s3://amzn-s3-demo-bucket/artifacts/file/tickets/" > now()
FileFreshness "s3://amzn-s3-demo-bucket/artifacts/file/tickets/" between (now() - 2
 hours) and (now() + 15 minutes)
FileFreshness "s3://amzn-s3-demo-bucket/artifacts/file/tickets/" between (now() - 3
 days) and (now() + 15 minutes)
FileFreshness "s3://amzn-s3-demo-bucket/artifacts/file/tickets/" between "2001-02-07"
 and (now() + 15 minutes)
FileFreshness "s3://amzn-s3-demo-bucket/artifacts/file/tickets/" > "21:45""
FileFreshness "s3://amzn-s3-demo-bucket/artifacts/file/tickets/" > "2024-01-01"
FileFreshness "s3://amzn-s3-demo-bucket/artifacts/file/tickets/" between "02:30" and
 "04:30"
FileFreshness "s3://amzn-s3-demo-bucket/artifacts/file/tickets/" between "9:30 AM" and
 "22:15"
```

Key considerations:

- FileFreshness can evaluate files using days, hours, and minute units
- For times, it supports AM / PM and 24-hour
- Times are calculated in UTC unless an override is specified
- Dates are calculated in UTC at time 00:00

FileFreshness that are time-based works as follows:

```
FileFreshness "s3://amzn-s3-demo-bucket/artifacts/file/tickets/" > "21:45"
```

- First, the time “21:45” is combined with today’s date in UTC format to create a date-time field
- Next, the date-time is converted to a timezone that you have specified
- Finally, the rule is evaluated

Optional File-based Rule Tags:

Tags allow you to control the rule behavior.

recentFiles

This tag limits the number of files processed by keeping the most recent file first.

```
FileFreshness "s3://amzn-s3-demo-bucket/" between (now() - 100 minutes) and (now() + 10
 minutes) with recentFiles = 1
```

timeZone

Accepted time zone overrides, see Allowed Time Zones for supported time zones.

```
FileFreshness "s3://path/" > "21:45" with timeZone = "America/New_York"
```

```
FileFreshness "s3://path/" > "21:45" with timeZone = "America/Chicago"
```

```
FileFreshness "s3://path/" > "21:45" with timeZone = "Europe/Paris"
```

```
FileFreshness "s3://path/" > "21:45" with timeZone = "Asia/Shanghai"
```

```
FileFreshness "s3://path/" > "21:45" with timeZone = "Australia/Darwin"
```

Inferring file names directly from data frames

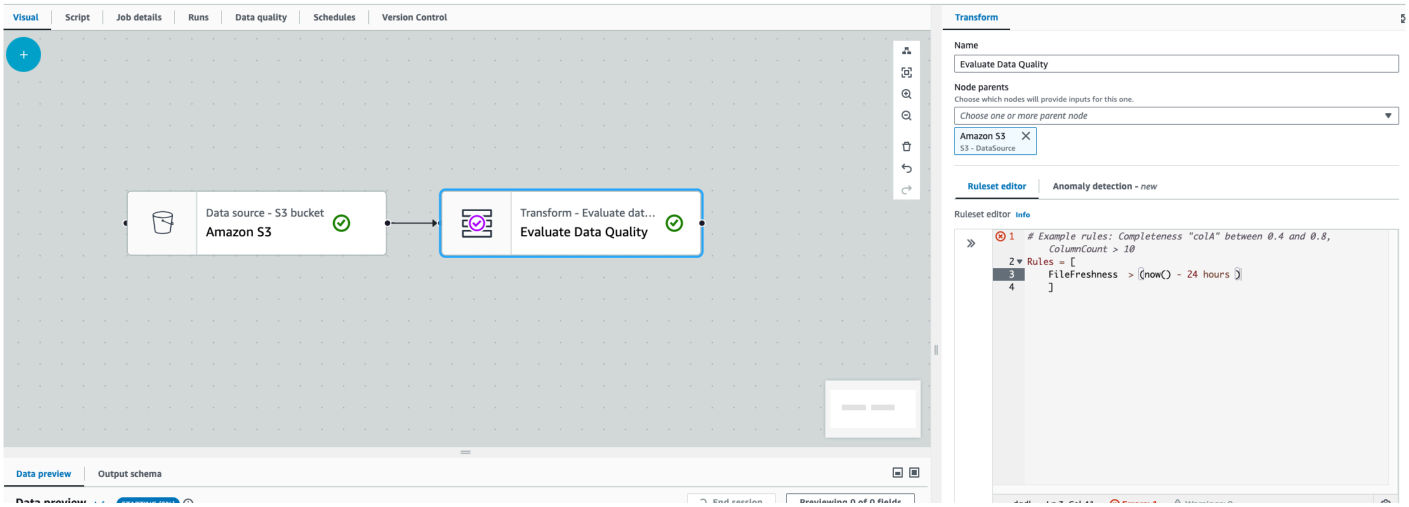
You don't always have to provide a file path. For instance, when you are authoring the rule in the AWS Glue Data Catalog, it may be hard to find which folders the catalog tables are using. AWS Glue Data Quality can find the specific folders or files used to populate your dataframe and can detect if they are fresh.

> **Note**
>
> This feature will only work when files are successfully read into the DynamicFrame or DataFrame.

```
FileFreshness > (now() - 24 hours)
```

This rule will find the folder path or files that are used to populate the dynamic frame or data frame. This works for Amazon S3 paths or Amazon S3-based AWS Glue Data Catalog tables. There are a few considerations:

1. In AWS Glue ETL, you must have the **EvaluateDataQuality** Transform immediately after an Amazon S3 or AWS Glue Data Catalog transform.


Visual
Script
Job details
Runs
Data quality
Schedules
Version Control
Data source - S3 bucket
Amazon S3
Transform - Evaluate dat...
Evaluate Data Quality
Data preview
Output schema
Transform
Ruleset editor
Anomaly detection - new
Info

Name: Evaluate Data Quality

Node parents: Amazon S3 (S3 - DataSource)

Ruleset editor

```
# Example rules: Completeness "colA" between 0.4 and 0.8,
    ColumnCount > 10
Rules = [
    FileFreshness  > (now() - 24 hours )
    ]
```

2. This rule will not work in AWS Glue Interactive Sessions.

If you attempt this in both of the cases, or when AWS Glue can’t find the files, AWS Glue will throw the following error: “`Unable to parse file path from DataFrame`”

FileMatch

The FileMatch rule allows you to compare files against other files or checksums. This can be useful in a few scenarios:

1. Validating files received from external sources: You can use FileMatch to ensure that you have received the correct files from external sources by comparing against checksums. This helps validate the integrity of the data you're ingesting.
2. Comparing data in two different folders: FileMatch can be used to compare files between two folders.

This rule gathers one metric: the number of files that were scanned by the rule.

```
{"Dataset.*.FileCount":1}
```

Validate file with a checksum:

FileMatch accepts a file and a set checksums to ensure that at least one checksums match the file.

```
FileMatch "s3://amzn-s3-demo-bucket/file.json" in ["3ee0d8617ac041793154713e5ef8f319"]
 with hashAlgorithm = "MD5"
FileMatch "s3://amzn-s3-demo-bucket/file.json" in ["3ee0d8617ac041793154713e5ef8f319"]
 with hashAlgorithm = "SHA-1"
FileMatch "s3://amzn-s3-demo-bucket/file.json" in ["3ee0d8617ac041793154713e5ef8f319"]
 with hashAlgorithm = "SHA-256"
FileMatch "s3://amzn-s3-demo-bucket/file.json" in ["3ee0d8617ac041793154713e5ef8f319"]
```

The following standard algorithms are supported:

- MD5
- SHA-1
- SHA-256

If you do not supply an algorithm, the default is SHA-256.

Validate all files in a folder with set of checksum:

```
FileMatch "s3://amzn-s3-demo-bucket /" in ["3ee0d8617ac041793154713e5ef8f319",
 "7e8617ac041793154713e5ef8f319"] with hashAlgorithm = "MD5"
FileMatch "s3://amzn-s3-demo-bucket /internal-folder/" in
 ["3ee0d8617ac041793154713e5ef8f319", "7e8617ac041793154713e5ef8f319"]
```

Compare files in different folders

```
FileMatch "s3://original_bucket/" "s3://archive_bucket/"
FileMatch "s3://original_bucket/internal-folder/" "s3://original_bucket/other-folder/"
```

FileMatch will check the contents of the files in `original_bucket` and ensure they match what’s in `archive_bucket`. The rule will fail if they don’t exactly match. It also can check the contents of internal folders or individual files.

FileMatch can also check individual files against each other.

```
FileMatch "s3://amzn-s3-demo-bucket /file_old.json" "s3://amzn-s3-demo-bucket /
file_new.json"
```

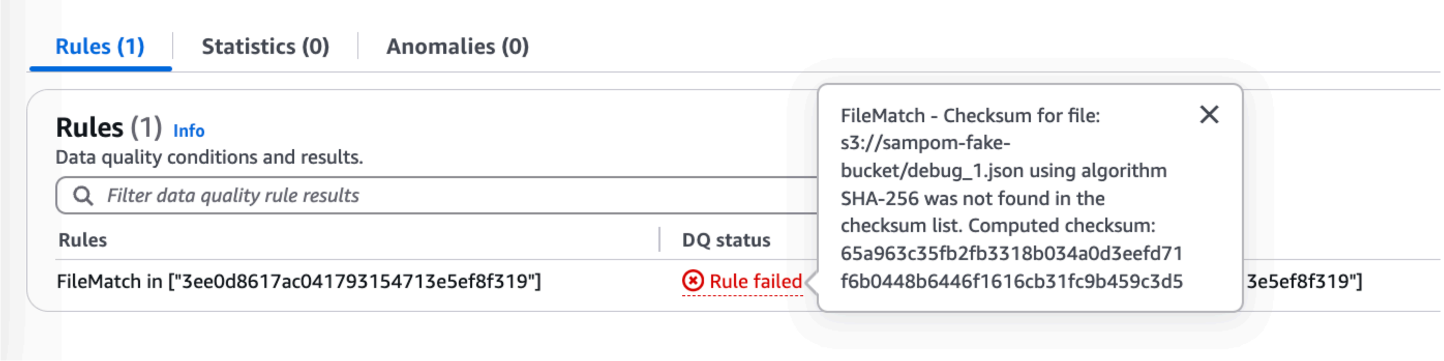
Inferring file names directly from data frames

You don't always have to provide a file path. For instance, when you are authoring the rule in the AWS Glue Data Catalog (backed by Amazon S3), it may be hard to find which folders the catalog tables are using. AWS Glue Data Quality can find the specific folders or files used to populate your data frame.

> **Note**
>
> This feature will only work when files are successfully read into the DynamicFrame or DataFrame.

```
FileMatch in ["3ee0d8617ac041793154713e5ef8f319"] with hashAlgorithm = "MD5"
FileMatch in ["3ee0d8617ac041793154713e5ef8f319"] with hashAlgorithm = "SHA-1"
FileMatch in ["3ee0d8617ac041793154713e5ef8f319"] with hashAlgorithm = "SHA-256"
FileMatch in ["3ee0d8617ac041793154713e5ef8f319"]
```

If the supplied checksum is different than what's computed, FileMatch will alert you to the difference.



Data quality conditions and results.

Rules	DQ status
FileMatch in ["3ee0d8617ac041793154713e5ef8f319"]	Rule failed

FileMatch - Checksum for file: s3://sampom-fake-bucket/debug_1.json using algorithm SHA-256 was not found in the checksum list. Computed checksum: 65a963c35fb2fb3318b034a0d3eefd71f6b0448b6446f1616cb31fc9b459c3d5

Optional File-based Rule Tags:

Tags allow you to control the rule behaviour.

recentFiles

This tag limits the number of files processed by keeping the most recent file first.

```
FileMatch "s3://amzn-s3-demo-bucket/file.json" in ["3ee0d8617ac04179sam4713e5ef8f319"]
 with recentFiles = 1
```

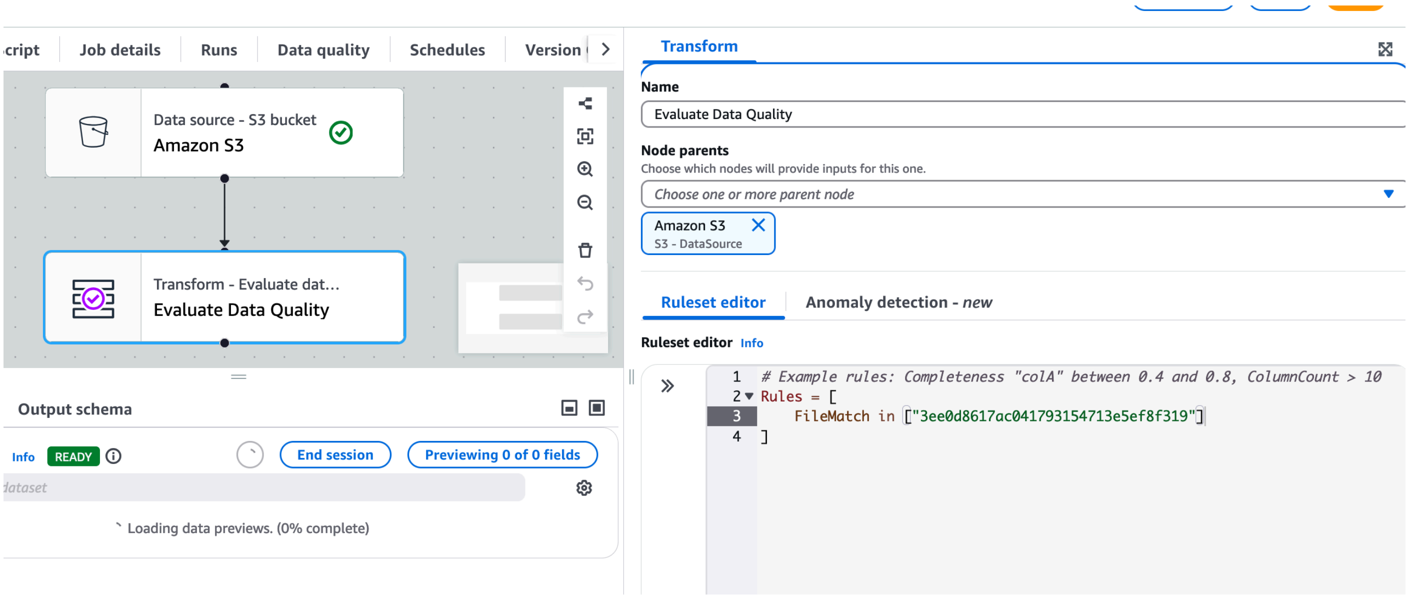
matchFileName

This tag ensures that files don't have duplicate names. Default behavior is false.

```
FileMatch "s3://amzn-s3-demo-bucket/file.json" in ["3ee0d8617ac04179sam4713e5ef8f319"]
 with matchFileName = "true"
```

There are a few considerations:

1. In AWS Glue ETL, you must have the **EvaluateDataQuality** Transform immediately after an Amazon S3 or AWS Glue Data Catalog transform.


cript
Job details
Runs
Data quality
Schedules
Version
Data source - S3 bucket
Amazon S3
Transform - Evaluate dat...
Evaluate Data Quality
Output schema
Info
READY
End session
Previewing 0 of 0 fields
dataset
Loading data previews. (0% complete)
Transform
Ruleset editor
Anomaly detection - new
Ruleset editor
Info

Name: Evaluate Data Quality

Node parents: Choose which nodes will provide inputs for this one.

Choose one or more parent node

Amazon S3 – S3 - DataSource

```
# Example rules: Completeness "colA" between 0.4 and 0.8, ColumnCount > 10
Rules = [
    FileMatch in ["3ee0d8617ac041793154713e5ef8f319"]
]
```

2. This rule will not work in AWS Glue Interactive Sessions.

FileUniqueness

File Uniqueness allows you to ensure that there are no duplicate files in the data you have received from your data producers.

It gathers the following data statistics:

1. The number of files that were scanned by the rule
2. The Uniqueness Ratio of the files

```
Dataset.*.FileUniquenessRatio: 1.00, Dataset.*.FileCount: 8.00
```

Find duplicate files in a folder:

```
FileUniqueness "s3://bucket/" > 0.5
FileUniqueness "s3://bucket/folder/" = 1
```

Inferring folder names directly from data frames to detect duplicates:

You don't always have to provide a file path. For instance, when you are authoring the rule in the AWS Glue Data Catalog, it may be hard to find which folders the catalog tables are using. AWS Glue Data Quality can find the specific folders or files used to populate your data frame.

> **Note**
>
> When using inference, file-based rules can only detect files successfully read into the DynamicFrame or DataFrame.

```
FileUniqueness > 0.5
```

Optional File-based Rule Tags:

Tags allow you to control the rule behaviour.

recentFiles

This tag limits the number of files processed by keeping the most recent file first.

```
FileUniqueness "s3://amzn-s3-demo-bucket/" > 0.5 with recentFiles = 1
```

matchFileName

This tag ensures that files don’t have duplicate names. Default behavior is false.

```
FileUniqueness "s3://amzn-s3-demo-bucket/" > 0.5 with matchFileName = "true"
```

There are a few considerations:

1. In AWS Glue ETL, you must have the **EvaluateDataQuality** Transform immediately after an Amazon S3 or AWS Glue Data Catalog transform.
2. This rule will not work in AWS Glue Interactive Sessions.

FileSize

The FileSize ruletype allows you to ensure that files meet a certain file size criteria. This is useful for following use cases:

1. Ensure that producers are not sending empty or substantially smaller files for processing.
2. Ensure that your target buckets don't have smaller files which may lead to performance issues.

FileSize gathers the following metrics:

1. Compliance: returns the % of files that meet the rule threshold you have established
2. File Count: the number of files that were scanned by the rule
3. Minimum file size in bytes
4. Maximum file size in bytes

```
Dataset.*.FileSize.Compliance: 1.00,
Dataset.*.FileCount: 8.00,
Dataset.*.MaximumFileSize: 327413121.00,
Dataset.*.MinimumFileSize: 204558920.00
```

Anomaly detection is not supported for these metrics.

Validate size of files

This rule will pass when file.dat is greater than 2 MB.

```
FileSize "s3://amzn-s3-demo-bucket/file.dat" > 2 MB
```

The supported unites include B(bytes), MB(mega bytes), GB(giga bytes) and TB(terra bytes).

Validate size of files in folders

```

FileSize "s3://bucket/" > 5 B
FileSize "s3://bucket/" < 2 GB
```

This rule will pass if 70% of the files in s3://amzn-s3-demo-bucket is between 2 GB and 1 TB.

```
FileSize "s3://amzn-s3-demo-bucket/" between 2 GB and 1 TB  with threshold > 0.7
```

Inferring file names directly from data frames

You don't always have to provide a file path. For instance, when you are authoring the rule in the Data Catalog, it may be hard to find which folders the catalog tables are using. AWS Glue Data Quality can find the specific folders or files used to populate your data frame.

> **Note**
>
> This feature will only work when files are successfully read into the DynamicFrame or DataFrame.

```
FileSize < 10 MB with threshold > 0.7
```

Optional File-based Rule Tags:

Tags allow you to control the rule behaviour.

recentFiles

This tag limits the number of files processed by keeping the most recent file first.

```
FileSize "s3://amzn-s3-demo-bucket/" > 5 B with recentFiles = 1
```

matchFileName

This tag ensures that files don’t have duplicate names. Default behavior is false.

```
FileSize "s3://amzn-s3-demo-bucket/" > 5 B with matchFileName = "true"
```

There are a few considerations:

1. In AWS Glue ETL, you must have Evaluate DataQuality Transform immediately after the Amazon S3 or Data Catalog transform.
2. This rule will not work in AWS Glue Interactive Sessions.

Using APIs to measure and manage data quality

This topic describes how to use APIs to measure and manage data quality.

Contents

- Prerequisites
- Working with AWS Glue Data Quality recommendations
- Working with AWS Glue Data Quality rulesets
- Working with AWS Glue Data Quality runs
- Working with AWS Glue Data Quality results

Prerequisites

- Make sure your boto3 version is up to date so that it includes the latest AWS Glue Data Quality API.
- Make sure your AWS CLI version is up to date, so as to include the latest CLI.

If you're using an AWS Glue job to run these APIs, you can use the following option to update the boto3 library to the latest version:

```
—additional-python-modules boto3==<version>
```

Working with AWS Glue Data Quality recommendations

To start an AWS Glue Data Quality recommendation run:

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""
    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 AWS Glue client.
        """
        self.glue_client = glue_client

    def start_data_quality_rule_recommendation_run(self, database_name, table_name,
 role_arn):
        """
        Starts a recommendation run that is used to generate rules when you don't know
 what rules to write. AWS Glue Data Quality
        analyzes the data and comes up with recommendations for a potential ruleset.
 You can then triage the ruleset
        and modify the generated ruleset to your liking.

        :param database_name: The name of the AWS Glue database which contains the
 dataset.
        :param table_name: The name of the AWS Glue table against which we want a
 recommendation
        :param role_arn: The Amazon Resource Name (ARN) of an AWS Identity and Access
 Management (IAM) role that grants permission to let AWS Glue access the resources it
 needs.

        """
        try:
            response = self.client.start_data_quality_rule_recommendation_run(
                DataSource={
                    'GlueTable': {
                        'DatabaseName': database_name,
                        'TableName': table_name
                    }
                },
                Role=role_arn
            )
        except ClientError as err:
            logger.error(
                "Couldn't start data quality recommendation run %s. Here's why: %s:
 %s", name,
                err.response['Error']['Code'], err.response['Error']['Message'])
            raise
        else:
            return response['RunId']
```

For a recommendation run, you are able to use your `pushDownPredicates` or `catalogPartitionPredicates` to improve performance and run recommendations only on specific partitions of your catalog sources.

```
client.start_data_quality_rule_recommendation_run(
            DataSource={
                'GlueTable': {
                    'DatabaseName': database_name,
                    'TableName': table_name,
                    'AdditionalOptions': {
                        'pushDownPredicate': "year=2022"
                    }
                }
            },
            Role=role_arn,
            NumberOfWorkers=2,
            CreatedRulesetName='<rule_set_name>'
  )
```

To get results of an AWS Glue Data Quality recommendation run:

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""
    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 AWS Glue client.
        """
        self.glue_client = glue_client

    def get_data_quality_rule_recommendation_run(self, run_id):
        """
        Gets the specified recommendation run that was used to generate rules.

        :param run_id: The id of the data quality recommendation run

        """
        try:
            response =
 self.client.get_data_quality_rule_recommendation_run(RunId=run_id)
        except ClientError as err:
            logger.error(
                "Couldn't get data quality recommendation run %. Here's why: %s: %s",
 run_id,
                err.response['Error']['Code'], err.response['Error']['Message'])
            raise
        else:
            return response
```

From the above response object, you can extract the RuleSet that was recommended by the run, to use in further steps:

```
print(response['RecommendedRuleset'])
```

```
Rules = [
    RowCount between 2000 and 8000,
    IsComplete "col1",
    IsComplete "col2",
    StandardDeviation "col3" between 58138330.8 and 64258155.09,
    ColumnValues "col4" between 1000042965 and 1214474826,
    IsComplete "col5"
]
```

To get a list of all your recommendation runs that can be filtered and listed:

```
response = client.list_data_quality_rule_recommendation_runs(
    Filter={
        'DataSource': {
            'GlueTable': {
                'DatabaseName': '<database_name>',
                'TableName': '<table_name>'
            }
        }
)
```

To cancel existing AWS Glue Data Quality recommendation tasks:

```
response = client.cancel_data_quality_rule_recommendation_run(
    RunId='dqrun-d4b6b01957fdd79e59866365bf9cb0e40fxxxxxxx'
)
```

Working with AWS Glue Data Quality rulesets

To create an AWS Glue Data Quality ruleset:

```
response = client.create_data_quality_ruleset(
    Name='<ruleset_name>',
    Ruleset='Rules = [IsComplete "col1", IsPrimaryKey "col2", RowCount between 2000 and
 8000]',
    TargetTable={
        'TableName': '<table_name>',
        'DatabaseName': '<database_name>'
    }
)
```

To get a data quality ruleset:

```
response = client.get_data_quality_ruleset(
    Name='<ruleset_name>'
)
print(response)
```

You can use this API to then extract the rule set:

```
print(response['Ruleset'])
```

To list all the data quality rulesets for a table:

```
response = client.list_data_quality_rulesets()
```

You can use the filter condition within the API to filter all rulesets attached to a specific database or table:

```
response = client.list_data_quality_rulesets(
    Filter={
        'TargetTable': {
            'TableName': '<table_name>',
            'DatabaseName': '<database_name>'
        }
    },
)
```

To update a data quality ruleset:

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""
    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 AWS Glue client.
        """
        self.glue_client = glue_client

    def update_data_quality_ruleset(self, ruleset_name, ruleset_string):
        """
        Update an AWS Glue Data Quality Ruleset
        :param ruleset_name: The name of the AWS Glue Data Quality ruleset to update
        :param ruleset_string: The DQDL ruleset string to update the ruleset with

        """
        try:
            response = self.client.update_data_quality_ruleset(
                Name=ruleset_name,
                Ruleset=ruleset_string
            )
        except ClientError as err:
            logger.error(
                "Couldn't update the AWS Glue Data Quality ruleset. Here's why: %s:
 %s",
                err.response['Error']['Code'], err.response['Error']['Message'])
            raise
        else:
            return response
```

To delete a data quality ruleset:

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""
    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 AWS Glue client.
        """
        self.glue_client = glue_client

    def delete_data_quality_ruleset(self, ruleset_name):
        """
        Delete a AWS Glue Data Quality Ruleset

        :param ruleset_name: The name of the AWS Glue Data Quality ruleset to delete

        """
        try:
            response = self.client.delete_data_quality_ruleset(
                Name=ruleset_name
            )
        except ClientError as err:
            logger.error(
                "Couldn't delete the AWS Glue Data Quality ruleset. Here's why: %s:
 %s",
                err.response['Error']['Code'], err.response['Error']['Message'])
            raise
        else:
            return response
```

Working with AWS Glue Data Quality runs

To start an AWS Glue Data Quality run:

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""
    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 AWS Glue client.
        """
        self.glue_client = glue_client

    def start_data_quality_ruleset_evaluation_run(self, database_name, table_name,
 role_name, ruleset_list):
        """
        Start an AWS Glue Data Quality evaluation run

        :param database_name: The name of the AWS Glue database which contains the
 dataset.
        :param table_name: The name of the AWS Glue table against which we want to
 evaluate.
        :param role_arn: The Amazon Resource Name (ARN) of an AWS Identity and Access
 Management (IAM) role that grants permission to let AWS Glue access the resources it
 needs.
        :param ruleset_list: The list of AWS Glue Data Quality ruleset names to
 evaluate.

        """
        try:
            response = client.start_data_quality_ruleset_evaluation_run(
                DataSource={
                    'GlueTable': {
                        'DatabaseName': database_name,
                        'TableName': table_name
                    }
                },
                Role=role_name,
                RulesetNames=ruleset_list
            )
        except ClientError as err:
            logger.error(
                "Couldn't start the AWS Glue Data Quality Run. Here's why: %s: %s",
                err.response['Error']['Code'], err.response['Error']['Message'])
            raise
        else:
            return response['RunId']
```

Remember that you can pass a `pushDownPredicate` or `catalogPartitionPredicate` parameter to ensure your data quality run only targets a specific set of partition within your catalog table. For example:

```
response = client.start_data_quality_ruleset_evaluation_run(
    DataSource={
        'GlueTable': {
            'DatabaseName': '<database_name>',
            'TableName': '<table_name>',
            'AdditionalOptions': {
                'pushDownPredicate': 'year=2023'
            }
        }
    },
    Role='<role_name>',
    NumberOfWorkers=5,
    Timeout=123,
    AdditionalRunOptions={
        'CloudWatchMetricsEnabled': False
    },
    RulesetNames=[
        '<ruleset_name>',
    ]
)
```

You can also configure how composite rules in your ruleset are evaluated, either at the ROW or COLUMN level. For more information on how composite rules work, please refer to How composite rules work in the documentation.

Example on how to set the composite rule evaluation method in your request:

```
response = client.start_data_quality_ruleset_evaluation_run(
    DataSource={
        'GlueTable': {
            'DatabaseName': '<database_name>',
            'TableName': '<table_name>',
            'AdditionalOptions': {
                'pushDownPredicate': 'year=2023'
            }
        }
    },
    Role='<role_name>',
    NumberOfWorkers=5,
    Timeout=123,
    AdditionalRunOptions={
        'CompositeRuleEvaluationMethod':ROW
    },
    RulesetNames=[
        '<ruleset_name>',
    ]
)
```

To get information about an AWS Glue Data Quality run:

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""
    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 AWS Glue client.
        """
        self.glue_client = glue_client

    def get_data_quality_ruleset_evaluation_run(self, run_id):
        """
        Get details about an AWS Glue Data Quality Run

        :param run_id: The AWS Glue Data Quality run ID to look up

        """
        try:
            response = self.client.get_data_quality_ruleset_evaluation_run(
                RunId=run_id
            )
        except ClientError as err:
            logger.error(
                "Couldn't look up the AWS Glue Data Quality run ID. Here's why: %s:
 %s",
                err.response['Error']['Code'], err.response['Error']['Message'])
            raise
        else:
            return response
```

To get the results from an AWS Glue Data Quality run:

For a given AWS Glue Data Quality run, you can extract the results of the run's evaluation using the following method:

```
response = client.get_data_quality_ruleset_evaluation_run(
    RunId='d4b6b01957fdd79e59866365bf9cb0e40fxxxxxxx'
)

resultID = response['ResultIds'][0]

response = client.get_data_quality_result(
    ResultId=resultID
)

print(response['RuleResults'])
```

To list all your AWS Glue Data Quality runs:

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""
    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 AWS Glue client.
        """
        self.glue_client = glue_client

    def list_data_quality_ruleset_evaluation_runs(self, database_name, table_name):
        """
        Lists all the AWS Glue Data Quality runs against a given table

        :param database_name: The name of the database where the data quality runs
        :param table_name: The name of the table against which the data quality runs
 were created

        """
        try:
            response = self.client.list_data_quality_ruleset_evaluation_runs(
                Filter={
                    'DataSource': {
                        'GlueTable': {
                            'DatabaseName': database_name,
                            'TableName': table_name
                        }
                    }
                }
            )
        except ClientError as err:
            logger.error(
                "Couldn't list the AWS Glue Quality runs. Here's why: %s: %s",
                err.response['Error']['Code'], err.response['Error']['Message'])
            raise
        else:
            return response
```

You can modify the filter clause to only show results between specific times or running against specific tables.

To stop an ongoing AWS Glue Data Quality run:

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""
    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 AWS Glue client.
        """
        self.glue_client = glue_client

    def cancel_data_quality_ruleset_evaluation_run(self, result_id):
        """
        Cancels a given AWS Glue Data Quality run

        :param result_id: The result id of a AWS Glue Data Quality run to cancel

        """
        try:
            response = self.client.cancel_data_quality_ruleset_evaluation_run(
                ResultId=result_id
            )
        except ClientError as err:
            logger.error(
                "Couldn't cancel the AWS Glue Data Quality run. Here's why: %s: %s",
                err.response['Error']['Code'], err.response['Error']['Message'])
            raise
        else:
            return response
```

Working with AWS Glue Data Quality results

To get your AWS Glue Data Quality run results:

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""
    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 AWS Glue client.
        """
        self.glue_client = glue_client

    def get_data_quality_result(self, result_id):
        """
        Outputs the result of an AWS Glue Data Quality Result

        :param result_id: The result id of an AWS Glue Data Quality run

        """
        try:
            response = self.client.get_data_quality_result(
                ResultId=result_id
            )
        except ClientError as err:
            logger.error(
                "Couldn't get the AWS Glue Data Quality result. Here's why: %s: %s",
                err.response['Error']['Code'], err.response['Error']['Message'])
            raise
        else:
            return response
```

To view the statistics gathered for a given data quality result:

```
import boto3
```

```
from botocore.exceptions import ClientError
import logging

logger = logging.getLogger(__name__)
class GlueWrapper:
    """Encapsulates AWS Glue actions."""
    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 AWS Glue client.
        """
        self.glue_client = glue_client

    def get_profile_for_data_quality_result(self, result_id):
        """
        Outputs the statistic profile for a AWS Glue Data Quality Result

        :param result_id: The result id of a AWS Glue Data Quality run

        """
        try:
            response = self.glue_client.get_data_quality_result(
                ResultId=result_id
            )

            # the profile contains all statistics gathered for the result
            profile_id = response['ProfileId']
            profile = self.glue_client.list_data_quality_statistics(
                ProfileId = profile_id
            )
            return profile
        except ClientError as err:
            logger.error(
                "Couldn't retrieve Data Quality profile. Here's why: %s: %s",
                err.response['Error']['Code'], err.response['Error']['Message'])
            raise
```

To view the timeseries for a statistic gathered across multiple data quality runs:

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""
    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 AWS Glue client.
        """
        self.glue_client = glue_client

    def get_statistics_for_data_quality_result(self, profile_id):
        """
        Outputs an array of datapoints for each statistic in the input result.

        :param result_id: The profile id of a AWS Glue Data Quality run

        """
        try:
            profile = self.glue_client.list_data_quality_statistics(
                ProfileId = profile_id
            )
            statistics = [self.glue_client.list_data_quality_statistics(
                StatisticId = s['StatisticId']
            ) for s in profile['Statistics']]
            return statistics
        except ClientError as err:
            logger.error(
                "Couldn't retrieve Data Quality statistics. Here's why: %s: %s",
                err.response['Error']['Code'], err.response['Error']['Message'])
            raise
```

To view the anomaly detection model for a specific statistic:

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""
    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 AWS Glue client.
        """
        self.glue_client = glue_client

    def get_model_training_result_for_statistic(self, statistic_id, profile_id):
        """
        Outputs the details (bounds) of anomaly detection training for the given
 statistic at the given profile.

        :param statistic_id the model's statistic (the timeseries it is tracking)
        :param profile_id the profile associated with the model (a point in the
 timeseries)

```

```
        """
        try:
            model = self.glue_client.get_data_quality_model_result(
                ProfileId = profile_id, StatisticId = statistic_id
            )
            return model
        except ClientError as err:
            logger.error(
                "Couldn't retrieve Data Quality model results. Here's why: %s: %s",
                err.response['Error']['Code'], err.response['Error']['Message'])
            raise
```

To exclude a datapoint from its statistic model's anomaly detection baseline:

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""
    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 AWS Glue client.
        """
        self.glue_client = glue_client

    def apply_exclusions_to_statistic(self, statistic_id, profile_ids):
        """
        Annotate some points along a given statistic timeseries.

        This example excludes the provided values; INCLUDE can also be used to undo
 this action.

        :param statistic_id the statistic timeseries to annotate
        :param profile_id the profiles we want to exclude (points in the timeseries)

        """


        try:
            response = self.glue_client.batch_put_data_quality_statistic_annotation(
                    InclusionAnnotations = [
                        {'ProfileId': prof_id,
                        'StatisticId': statistic_id,
                        'InclusionAnnotation': 'EXCLUDE'} for prof_id in profile_ids
                    ]
            )
            return response['FailedInclusionAnnotations']
        except ClientError as err:
            logger.error(
                "Couldn't store Data Quality annotations. Here's why: %s: %s",
                err.response['Error']['Code'], err.response['Error']['Message'])
            raise
```

To view the status of anomaly detection model training for a specific statistic:

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""
    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 AWS Glue client.
        """
        self.glue_client = glue_client

    def get_model_training_status_for_statistic(self, statistic_id, profile_id):
        """
        Outputs the status of anomaly detection training for the given statistic at the
 given profile.

        :param statistic_id the model's statistic (the timeseries it is tracking)
        :param profile_id the profile associated with the model (a point in the
 timeseries)

        """
        try:
            model = self.glue_client.get_data_quality_model(
                ProfileId = profile_id, StatisticId = statistic_id
            )
            return model
        except ClientError as err:
            logger.error(
                "Couldn't retrieve Data Quality statistics. Here's why: %s: %s",
                err.response['Error']['Code'], err.response['Error']['Message'])
            raise
```

To exclude all results from a specific data quality run from anomaly detection baselines:

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""
    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 AWS Glue client.
        """
        self.glue_client = glue_client

    def apply_exclusions_to_profile(self, profile_id):
        """
        Exclude datapoints produced by a run across statistic timeseries.

        This example excludes the provided values; INCLUDE can also be used to undo
 this action.

        :param profile_id the profiles we want to exclude (points in the timeseries)

        """
        try:
            response = self.glue_client.put_data_quality_profile_annotation(
                    ProfileId = profile_id,
                    InclusionAnnotation = "EXCLUDE"
            )
            return response
        except ClientError as err:
            logger.error(
                "Couldn't store Data Quality annotations. Here's why: %s: %s",
                err.response['Error']['Code'], err.response['Error']['Message'])
            raise
```

To get the results from a given data quality run and display the results:

With a AWS Glue Data Quality `runID`, you can extract the `resultID` to then get the actual results, as shown below:

```
response = client.get_data_quality_ruleset_evaluation_run(
    RunId='dqrun-abca77ee126abe1378c1da1ae0750d7dxxxx'
)

resultID = response['ResultIds'][0]

response = client.get_data_quality_result(
    ResultId=resultID
)

print(resp['RuleResults'])
```

Setting up alerts, deployments, and scheduling

This topic describes how to set up alerts, deployments, and scheduling for AWS Glue Data Quality.

Contents

- Setting up alerts and notifications in Amazon EventBridge integration
  - Additional configuration options for the event pattern
  - Formatting notifications as emails
- Set up alerts and notifications in CloudWatch integration
- Querying data quality results to build dashboards
- Deploying data quality rules using AWS CloudFormation
- Scheduling data quality rules

Setting up alerts and notifications in Amazon EventBridge integration

AWS Glue Data Quality supports the publishing of EventBridge events, which are emitted upon completion of a Data Quality ruleset evaluation run. With this, you can easily setup alerts when data quality rules fail.

Here is a sample event when you evaluate data quality rulesets in the Data Catalog. With this information, you can review the data that is made available with Amazon EventBridge. You can issue additional API calls to get more details. For example, call the `get_data_quality_result` API with the result ID to get the details of a particular execution.

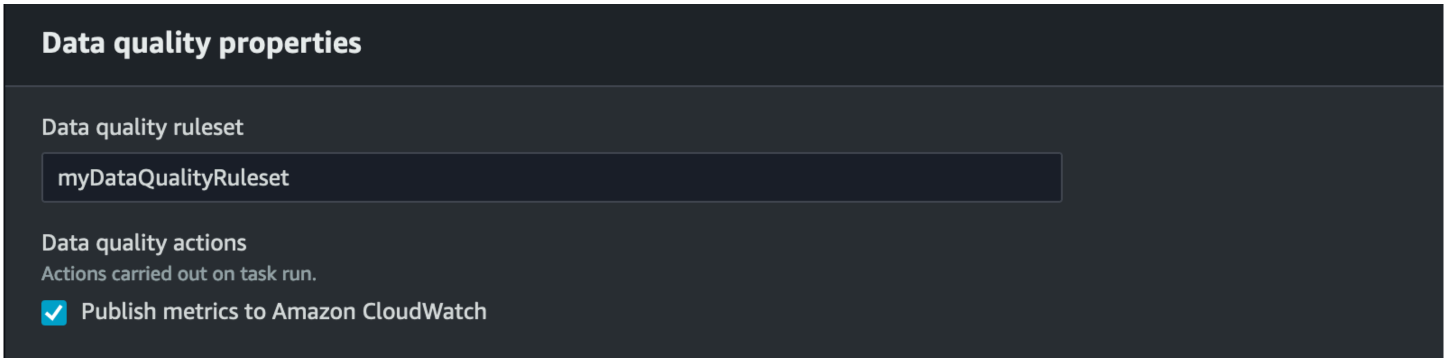
```
{
    "version":"0",
    "id":"abcdef00-1234-5678-9abc-def012345678",
    "detail-type":"Data Quality Evaluation Results Available",
    "source":"aws.glue-dataquality",
    "account":"123456789012",
    "time":"2017-09-07T18:57:21Z",
    "region":"us-west-2",
    "resources":[],
    "detail":{
        "context": {
                        "contextType": "GLUE_DATA_CATALOG",
                        "runId":"dqrun-12334567890",
                        "databaseName": "db-123",
                        "tableName": "table-123",
                        "catalogId": "123456789012"
                        },
            "resultID": "dqresult-12334567890",
            "rulesetNames": ["rulset1"],
            "state":"SUCCEEDED",
            "score": 1.00,
            "rulesSucceeded": 100,
            "rulesFailed": 0,
            "rulesSkipped": 0
    }
}
```

Here is a sample event that gets published when you evaluate data quality rulesets in AWS Glue ETL or AWS Glue Studio notebooks.

```
{
    "version":"0",
    "id":"abcdef00-1234-5678-9abc-def012345678",
    "detail-type":"Data Quality Evaluation Results Available",
    "source":"aws.glue-dataquality",
    "account":"123456789012",
    "time":"2017-09-07T18:57:21Z",
    "region":"us-west-2",
    "resources":[],
    "detail":{
        "context": {
                        "contextType": "GLUE_JOB",
                        "jobId": "jr-12334567890",
                        "jobName": "dq-eval-job-1234",
                        "evaluationContext": "",
                        }
        "resultID": "dqresult-12334567890",
        "rulesetNames": ["rulset1"],
        "state":"SUCCEEDED",
        "score": 1.00
        "rulesSucceeded": 100,
        "rulesFailed": 0,
        "rulesSkipped": 0
    }
}
```

For Data Quality evaluation runs both in the Data Catalog and in ETL jobs, the **Publish metrics to Amazon CloudWatch** option, which is selected by default, must remain selected for EventBridge publishing to work.

Setting up EventBridge notifications



Data quality properties

Data quality ruleset

myDataQualityRuleset

Data quality actions

Actions carried out on task run.

☑ Publish metrics to Amazon CloudWatch

To receive the emitted events and define targets, you must configure Amazon EventBridge rules. To create rules:

1. Open the Amazon EventBridge console.
2. Choose **Rules** under the **Buses** section of the navigation bar.
3. Choose **Create Rule**.
4. On **Define Rule Detail**:
   a. For Name, enter `myDQRule`.
   b. Enter the description (optional).
   c. For event bus, select your event bus. If you don't have one, leave it as default.
   d. For Rule type select **Rule with an event pattern** then choose **Next**.
5. On **Build Event Pattern**:
   a. For event source select **AWS events or EventBridge partner events**.
   b. Skip the sample event section.
   c. For creation method select **Use pattern form**.
   d. For event pattern:
      i. Select **AWS services** for Event source.
      ii. Select **Glue Data Quality** for AWS service.
      iii. Select **Data Quality Evaluation Results Available** for Event type.
      iv. Select **FAILED** for Specific state(s). Then you see an event pattern similar to the following:

```
{
  "source": ["aws.glue-dataquality"],
  "detail-type": ["Data Quality Evaluation Results Available"],
  "detail": {
    "state": ["FAILED"]
  }
}
```

   v. For more configuration options see Additional configuration options for the event pattern.

6. On **Select Target(s)**:
   a. For **Target Types** select **AWS service**.
   b. Use the **Select a target** dropdown to choose your desired AWS service to connect to (SNS, Lambda, SQS, etc.), then choose **Next**.
7. On **Configure tag(s)** click **Add new tags** to add optional tags then choose **Next**.
8. You see a summary page of all the selections. Choose **Create rule** at the bottom.

Additional configuration options for the event pattern

In addition to filtering your event on success or failure, you may want to further filter events on different parameters.

To do this, go to the Event Pattern section, and select **Edit pattern** to specify additional parameters. Note that fields in the event pattern are case sensitive. The following are examples of configuring the event pattern.

To capture events from a particular table evaluating specific rulesets use this type of pattern:

```
{
  "source": ["aws.glue-dataquality"],
  "detail-type": ["Data Quality Evaluation Results Available"],
  "detail": {
    "context": {
      "contextType": ["GLUE_DATA_CATALOG"],
      "databaseName": "db-123",
       "tableName": "table-123",
    },
    "rulesetNames": ["ruleset1", "ruleset2"]
    "state": ["FAILED"]
  }
}
```

To capture events from specific jobs in the ETL experience use this type of pattern:

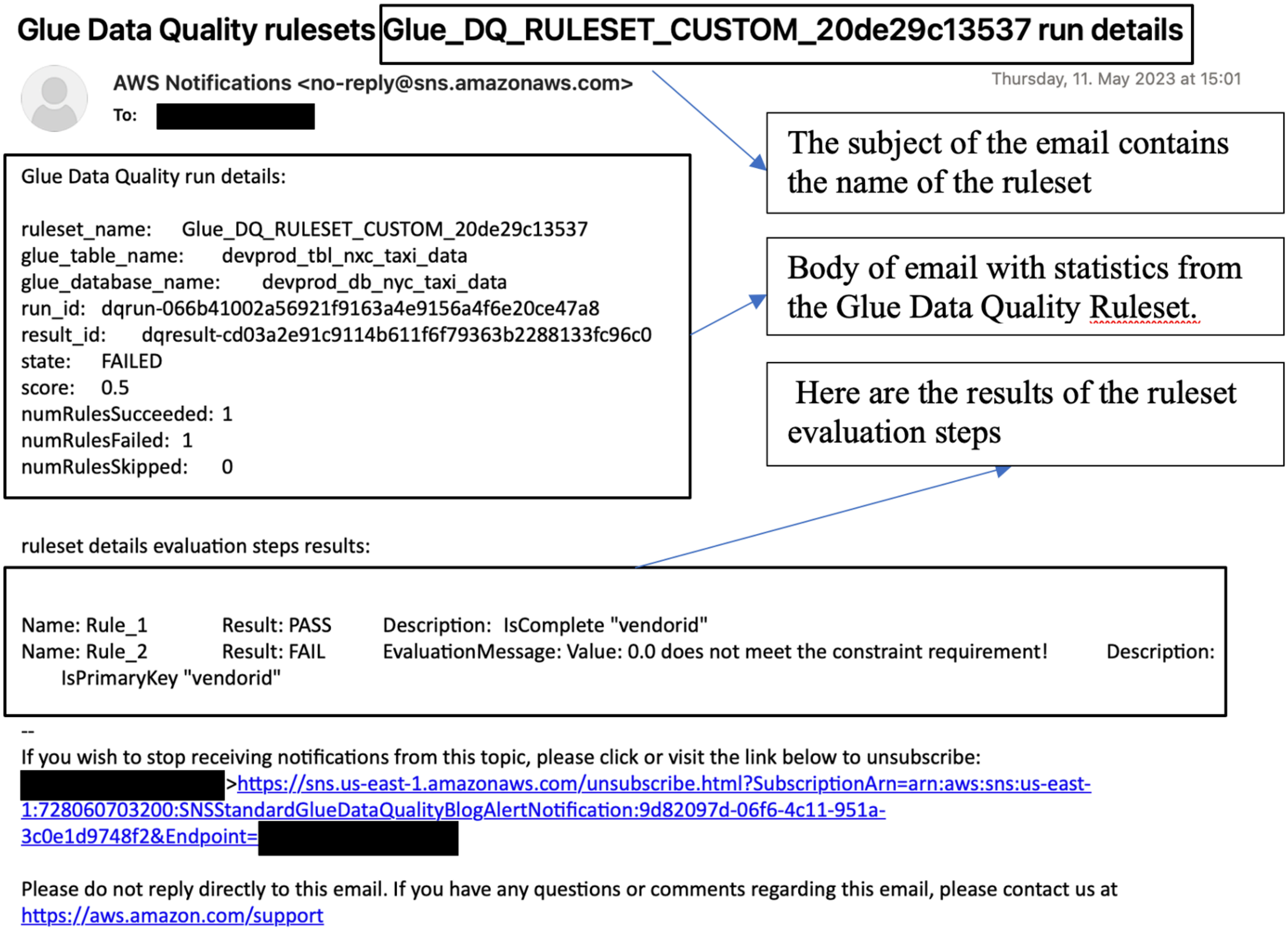
```
{
  "source": ["aws.glue-dataquality"],
  "detail-type": ["Data Quality Evaluation Results Available"],
  "detail": {
    "context": {
      "contextType": ["GLUE_JOB"],
      "jobName": ["dq_evaluation_job1", "dq_evaluation_job2"]
    },
    "state": ["FAILED"]
  }
}
```

To capture events with a score under a specific threshold (e.g. 70%):

```
{
  "source": ["aws.glue-dataquality"],
  "detail-type": ["Data Quality Evaluation Results Available"],
  "detail": {
    "score": [{
      "numeric": ["<=", 0.7]
    }]
  }
}
```

Formatting notifications as emails

Sometimes you need to send a well-formatted email notification to your business teams. You can use Amazon EventBridge and AWS Lambda to achieve this.



Glue Data Quality rulesets Glue_DQ_RULESET_CUSTOM_20de29c13537 run details

AWS Notifications <no-reply@sns.amazonaws.com>
To: [redacted]

Thursday, 11. May 2023 at 15:01

The subject of the email contains the name of the ruleset

Body of email with statistics from the Glue Data Quality Ruleset.

Here are the results of the ruleset evaluation steps

Glue Data Quality run details:

ruleset_name: Glue_DQ_RULESET_CUSTOM_20de29c13537
glue_table_name: devprod_tbl_nxc_taxi_data
glue_database_name: devprod_db_nyc_taxi_data
run_id: dqrun-066b41002a56921f9163a4e9156a4f6e20ce47a8
result_id: dqresult-cd03a2e91c9114b611f6f79363b2288133fc96c0
state: FAILED
score: 0.5
numRulesSucceeded: 1
numRulesFailed: 1
numRulesSkipped: 0

ruleset details evaluation steps results:

Name: Rule_1 Result: PASS Description: IsComplete "vendorid"
Name: Rule_2 Result: FAIL EvaluationMessage: Value: 0.0 does not meet the constraint requirement! Description: IsPrimaryKey "vendorid"

--
If you wish to stop receiving notifications from this topic, please click or visit the link below to unsubscribe:
[redacted]>https://sns.us-east-1.amazonaws.com/unsubscribe.html?SubscriptionArn=arn:aws:sns:us-east-1:728060703200:SNSStandardGlueDataQualityBlogAlertNotification:9d82097d-06f6-4c11-951a-3c0e1d9748f2&Endpoint=[redacted]

Please do not reply directly to this email. If you have any questions or comments regarding this email, please contact us at https://aws.amazon.com/support

The following sample code can be used to format your data quality notifications to generate emails.

```
import boto3
import json
from datetime import datetime

sns_client = boto3.client('sns')
glue_client = boto3.client('glue')

sns_topic_arn = 'arn:aws:sns:<region-code>:<account-id>:<sns-topic-name>'



def lambda_handler(event, context):
    log_metadata = {}
    message_text = ""
    subject_text = ""

    if event['detail']['context']['contextType'] == 'GLUE_DATA_CATALOG':
        log_metadata['ruleset_name'] = str(event['detail']['rulesetNames'][0])
        log_metadata['tableName'] = str(event['detail']['context']['tableName'])
        log_metadata['databaseName'] = str(event['detail']['context']['databaseName'])
        log_metadata['runId'] = str(event['detail']['context']['runId'])
        log_metadata['resultId'] = str(event['detail']['resultId'])
        log_metadata['state'] = str(event['detail']['state'])
        log_metadata['score'] = str(event['detail']['score'])
        log_metadata['numRulesSucceeded'] = str(event['detail']['numRulesSucceeded'])
        log_metadata['numRulesFailed'] = str(event['detail']['numRulesFailed'])
        log_metadata['numRulesSkipped'] = str(event['detail']['numRulesSkipped'])

        message_text += "Glue Data Quality run details:\n"
        message_text += "ruleset_name: {}\n".format(log_metadata['ruleset_name'])
        message_text += "glue_table_name: {}\n".format(log_metadata['tableName'])
        message_text += "glue_database_name: {}\n".format(log_metadata['databaseName'])
        message_text += "run_id: {}\n".format(log_metadata['runId'])
        message_text += "result_id: {}\n".format(log_metadata['resultId'])
        message_text += "state: {}\n".format(log_metadata['state'])
        message_text += "score: {}\n".format(log_metadata['score'])
        message_text += "numRulesSucceeded: 
{}\n".format(log_metadata['numRulesSucceeded'])
        message_text += "numRulesFailed: {}\n".format(log_metadata['numRulesFailed'])
        message_text += "numRulesSkipped: {}\n".format(log_metadata['numRulesSkipped'])

        subject_text = "Glue Data Quality ruleset {} run 
details".format(log_metadata['ruleset_name'])

    else:
        log_metadata['ruleset_name'] = str(event['detail']['rulesetNames'][0])
        log_metadata['jobName'] = str(event['detail']['context']['jobName'])
        log_metadata['jobId'] = str(event['detail']['context']['jobId'])
        log_metadata['resultId'] = str(event['detail']['resultId'])
        log_metadata['state'] = str(event['detail']['state'])
        log_metadata['score'] = str(event['detail']['score'])

        log_metadata['numRulesSucceeded'] = str(event['detail']['numRulesSucceeded'])
        log_metadata['numRulesFailed'] = str(event['detail']['numRulesFailed'])
        log_metadata['numRulesSkipped'] = str(event['detail']['numRulesSkipped'])

        message_text += "Glue Data Quality run details:\n"
        message_text += "ruleset_name: {}\n".format(log_metadata['ruleset_name'])
        message_text += "glue_job_name: {}\n".format(log_metadata['jobName'])
        message_text += "job_id: {}\n".format(log_metadata['jobId'])
        message_text += "result_id: {}\n".format(log_metadata['resultId'])
        message_text += "state: {}\n".format(log_metadata['state'])
        message_text += "score: {}\n".format(log_metadata['score'])
        message_text += "numRulesSucceeded: 
{}\n".format(log_metadata['numRulesSucceeded'])
        message_text += "numRulesFailed: {}\n".format(log_metadata['numRulesFailed'])
        message_text += "numRulesSkipped: {}\n".format(log_metadata['numRulesSkipped'])

        subject_text = "Glue Data Quality ruleset {} run 
details".format(log_metadata['ruleset_name'])

    resultID = str(event['detail']['resultId'])
    response = glue_client.get_data_quality_result(ResultId=resultID)
    RuleResults = response['RuleResults']
    message_text += "\n\nruleset details evaluation steps results:\n\n"
    subresult_info = []

    for dic in RuleResults:
        subresult = "Name: {}\t\tResult: {}\t\tDescription: \t{}".format(dic['Name'], 
dic['Result'], dic['Description'])
        if 'EvaluationMessage' in dic:
            subresult += "\t\tEvaluationMessage: {}".format(dic['EvaluationMessage'])
        subresult_info.append({
            'Name': dic['Name'],
            'Result': dic['Result'],
            'Description': dic['Description'],
            'EvaluationMessage': dic.get('EvaluationMessage', '')
        })
        message_text += "\n" + subresult

    log_metadata['resultrun'] = subresult_info



    sns_client.publish(
        TopicArn=sns_topic_arn,
        Message=message_text,
        Subject=subject_text
    )

    return {
        'statusCode': 200,
        'body': json.dumps('Message published to SNS topic')
    }
```

Set up alerts and notifications in CloudWatch integration

Our recommended approach is to set up data quality alerts using Amazon EventBridge, because Amazon EventBridge requires a one-time setup to alert customers. However, some customers prefer Amazon CloudWatch due to familiarity. For such customers, we offer integration with Amazon CloudWatch.

Each AWS Glue Data Quality evaluation emits a pair of metrics named `glue.data.quality.rules.passed` (indicating a number of rules that passed) and `glue.data.quality.rules.failed` (indicating the number of failed rules) per data quality run. You can use this emitted metric to create alarms to alert users if a given data quality run falls below a threshold. To get started with setting up an alarm that would send an email via an Amazon SNS notification, follow the steps below:

To get started with setting up an alarm that would send an email via an Amazon SNS notification, follow the steps below:

1. Open the Amazon CloudWatch console.
2. Choose **All metrics** under **Metrics**. You will see an additional namespace under Custom namespaces titled Glue Data Quality.

   > **Note**
   >
   > When starting an AWS Glue Data Quality run, make sure the **Publish metrics to Amazon CloudWatch** checkbox is enabled. Otherwise, metrics for that particular run will not be published to Amazon CloudWatch.

   Under the `Glue Data Quality` namespace, you can see metrics being emitted per table, per ruleset. For the purpose of this topic, we will use the `glue.data.quality.rules.failed` rule and alarm if this value goes over 1 (indicating that, if we see a number of failed rule evaluations greater than 1, we want to be notified).
3. To create the alarm, choose **All alarms** under **Alarms**.

4. Choose **Create alarm**.
5. Choose **Select metric**.
6. Select the `glue.data.quality.rules.failed` metric corresponding to the table you've created, then choose **Select metric**.
7. Under the **Specify metric and conditions** tab, under the **Metrics** section:
    a. For **Statistic**, choose **Sum**.
    b. For **Period**, choose **1 minute**.
8. Under the **Conditions** section:
    a. For **Threshold type**, choose **Static**.
    b. For **Whenever glue.data.quality.rules.failed is...**, select **Greater/Equal**.
    c. For **than...**, enter **1** as the threshold value.

    These selections imply that if the `glue.data.quality.rules.failed` metric emits a value greater than or equal to 1, we will trigger an alarm. However, if there is no data, we will treat it as acceptable.
9. Choose **Next**.
10. On **Configure actions**:
    a. For the **Alarm state trigger** section, choose **In alarm**.
    b. For **Send a notification to the following SNS topic** section, choose **Create a new topic to send a notification via a new SNS topic**.
    c. For **Email endpoints that will receive the notification** enter your email address. Then click **Create Topic**.
    d. Choose **Next**.
11. For **Alarm name**, enter `myFirstDQAlarm`, then choose **Next**.
12. You see a summary page of all the selections. Choose **Create alarm** at the bottom.

You can now see the alarm being created from the Amazon CloudWatch alarms dashboard.

Querying data quality results to build dashboards

You may want to build a dashboard to display your data quality results. There are two ways to do this:

Set up Amazon EventBridge with the following code to write the data to Amazon S3:

```
import boto3
import json
from datetime import datetime


s3_client = boto3.client('s3')
glue_client = boto3.client('glue')


s3_bucket = 's3-bucket-name'

def write_logs(log_metadata):
    try:
        filename = datetime.now().strftime("%m%d%Y%H%M%S") + ".json"
        key_opts = {
            'year': datetime.now().year,
            'month': "{:02d}".format(datetime.now().month),
            'day': "{:02d}".format(datetime.now().day),
            'filename': filename
        }
        s3key = "gluedataqualitylogs/year={year}/month={month}/day={day}/
{filename}".format(**key_opts)
        s3_client.put_object(Bucket=s3_bucket, Key=s3key,
 Body=json.dumps(log_metadata))
    except Exception as e:
        print(f'Error writing logs to S3: {e}')


def lambda_handler(event, context):
    log_metadata = {}
    message_text = ""
    subject_text = ""

    if event['detail']['context']['contextType'] == 'GLUE_DATA_CATALOG':
        log_metadata['ruleset_name'] = str(event['detail']['rulesetNames'][0])
        log_metadata['tableName'] = str(event['detail']['context']['tableName'])
        log_metadata['databaseName'] = str(event['detail']['context']['databaseName'])
        log_metadata['runId'] = str(event['detail']['context']['runId'])
        log_metadata['resultId'] = str(event['detail']['resultId'])
        log_metadata['state'] = str(event['detail']['state'])
        log_metadata['score'] = str(event['detail']['score'])
        log_metadata['numRulesSucceeded'] = str(event['detail']['numRulesSucceeded'])
        log_metadata['numRulesFailed'] = str(event['detail']['numRulesFailed'])
        log_metadata['numRulesSkipped'] = str(event['detail']['numRulesSkipped'])

        message_text += "Glue Data Quality run details:\n"
        message_text += "ruleset_name: {}\n".format(log_metadata['ruleset_name'])
        message_text += "glue_table_name: {}\n".format(log_metadata['tableName'])
        message_text += "glue_database_name: {}\n".format(log_metadata['databaseName'])
        message_text += "run_id: {}\n".format(log_metadata['runId'])
        message_text += "result_id: {}\n".format(log_metadata['resultId'])
        message_text += "state: {}\n".format(log_metadata['state'])
        message_text += "score: {}\n".format(log_metadata['score'])
        message_text += "numRulesSucceeded: 
{}\n".format(log_metadata['numRulesSucceeded'])
        message_text += "numRulesFailed: {}\n".format(log_metadata['numRulesFailed'])
        message_text += "numRulesSkipped: {}\n".format(log_metadata['numRulesSkipped'])

        subject_text = "Glue Data Quality ruleset {} run 
details".format(log_metadata['ruleset_name'])

    else:
        log_metadata['ruleset_name'] = str(event['detail']['rulesetNames'][0])
        log_metadata['jobName'] = str(event['detail']['context']['jobName'])
        log_metadata['jobId'] = str(event['detail']['context']['jobId'])
        log_metadata['resultId'] = str(event['detail']['resultId'])
        log_metadata['state'] = str(event['detail']['state'])
        log_metadata['score'] = str(event['detail']['score'])

        log_metadata['numRulesSucceeded'] = str(event['detail']['numRulesSucceeded'])
        log_metadata['numRulesFailed'] = str(event['detail']['numRulesFailed'])
        log_metadata['numRulesSkipped'] = str(event['detail']['numRulesSkipped'])

        message_text += "Glue Data Quality run details:\n"
        message_text += "ruleset_name: {}\n".format(log_metadata['ruleset_name'])
        message_text += "glue_job_name: {}\n".format(log_metadata['jobName'])
        message_text += "job_id: {}\n".format(log_metadata['jobId'])
        message_text += "result_id: {}\n".format(log_metadata['resultId'])
        message_text += "state: {}\n".format(log_metadata['state'])
        message_text += "score: {}\n".format(log_metadata['score'])
        message_text += "numRulesSucceeded: 
{}\n".format(log_metadata['numRulesSucceeded'])
        message_text += "numRulesFailed: {}\n".format(log_metadata['numRulesFailed'])
        message_text += "numRulesSkipped: {}\n".format(log_metadata['numRulesSkipped'])

        subject_text = "Glue Data Quality ruleset {} run 
details".format(log_metadata['ruleset_name'])
```

```
    resultID = str(event['detail']['resultId'])
    response = glue_client.get_data_quality_result(ResultId=resultID)
    RuleResults = response['RuleResults']
    message_text += "\n\nruleset details evaluation steps results:\n\n"
    subresult_info = []

    for dic in RuleResults:
        subresult = "Name: {}\t\tResult: {}\t\tDescription: \t{}".format(dic['Name'], 
 dic['Result'], dic['Description'])
        if 'EvaluationMessage' in dic:
            subresult += "\t\tEvaluationMessage: {}".format(dic['EvaluationMessage'])
        subresult_info.append({
            'Name': dic['Name'],
            'Result': dic['Result'],
            'Description': dic['Description'],
            'EvaluationMessage': dic.get('EvaluationMessage', '')
        })
        message_text += "\n" + subresult

    log_metadata['resultrun'] = subresult_info

    write_logs(log_metadata)


    return {
        'statusCode': 200,
        'body': json.dumps('Message published to SNS topic')
    }
```

After writing to Amazon S3, you can use AWS Glue crawlers to register to Athena and query the tables.

Configure an Amazon S3 location during a data quality evaluation::

When running data quality tasks in the AWS Glue Data Catalog or AWS Glue ETL, you can provide an Amazon S3 location to write the data quality results to Amazon S3. You can use the syntax below to create a table by referencing the target to read the data quality results.

Note that you must run the `CREATE EXTERNAL TABLE` and `MSCK REPAIR TABLE` queries separately.

```
CREATE EXTERNAL TABLE <my_table_name>(
    catalogid string,
    databasename string,
    tablename string,
    dqrunid string,
    evaluationstartedon timestamp,
    evaluationcompletedon timestamp,
    rule string,
    outcome string,
    failurereason string,
    evaluatedmetrics string)
PARTITIONED BY (
    `year` string,
    `month` string,
    `day` string)
ROW FORMAT SERDE 'org.openx.data.jsonserde.JsonSerDe'
WITH SERDEPROPERTIES (

 'paths'='catalogId,databaseName,dqRunId,evaluatedMetrics,evaluationCompletedOn,evaluationStart
STORED AS INPUTFORMAT 'org.apache.hadoop.mapred.TextInputFormat'
OUTPUTFORMAT 'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'
LOCATION 's3://glue-s3-dq-bucket-us-east-2-results/'
TBLPROPERTIES (
    'classification'='json',
    'compressionType'='none',
    'typeOfData'='file');
```

```
MSCK REPAIR TABLE <my_table_name>;
```

Once you create the above table, you can run analytical queries using Amazon Athena.

Deploying data quality rules using AWS CloudFormation

You can use AWS CloudFormation to create data quality rules. For more information, see AWS CloudFormation for AWS Glue.

Scheduling data quality rules

You can schedule data quality rules using the following methods:

- Schedule data quality rules from the Data Catalog: no code users can use this option to easily schedule their data quality scans. AWS Glue Data Quality will create the schedule in Amazon EventBridge. To schedule data quality rules:
  - Navigate to the ruleset and click **Run**.
  - In the **Run frequency**, select the desired schedule and provide a **Task Name**. This Task Name is the name of your schedule in EventBridge.
- Use Amazon EventBridge and AWS Step Functions to orchestrate evaluations and recommendations for data quality rules.

Data encryption at rest for AWS Glue Data Quality

AWS Glue Data Quality provides encryption by default to protect sensitive customer data at rest using AWS owned encryption keys.

AWS owned keys

AWS Glue Data Quality uses these keys to automatically encrypt customers' Data Quality assets. You cannot view, manage, or use AWS owned keys, or audit their use. However, you don't need to take any action or change any programs to protect the keys that encrypt your data. For more information, see AWS owned keys in the AWS KMS Developer Guide.

Encryption of data at rest by default helps reduce the operational overhead and complexity involved in protecting sensitive data. At the same time, it enables you to build secure applications that meet strict encryption compliance and regulatory requirements.

While you can't disable this layer of encryption or select an alternate encryption type, you can add a second layer of encryption over the existing AWS owned encryption keys by choosing a customer-managed key when you create your Data Quality resources.

Customer managed keys

Customer managed keys: AWS Glue Data Quality supports the use of a symmetric customer-managed key that you create, own, and manage. This adds a second layer of encryption over the existing AWS owned encryption. Because you have full control of this layer of encryption, you can perform tasks such as:

- Establishing and maintaining key policies

- Establishing and maintaining IAM policies
- Enabling and disabling key policies
- Rotating key cryptographic material
- Adding tags
- Creating key aliases
- Scheduling keys for deletion

For more information, see Customer managed keys in the AWS KMS Developer Guide.

The following table summarizes how AWS Glue Data Quality encrypts different Data Quality assets.

Data type	AWS owned key encryption	Customer managed key encryption
Data Quality Ruleset DQDL ruleset string that is referenced by the persisted DQ ruleset. These persisted rulesets are only used in the AWS Glue Data Catalog experience for now.	Enabled	Enabled
Data Quality Rule/Analyzer Results Result artifacts that contains the pass/fail status of each rule in a ruleset as well as metrics collected by both rules and analyzers.	Enabled	Enabled
Observations Observations are generated when an anomaly is detected in the data. It contains	Enabled	Enabled

Data type	AWS owned key encryption	Customer managed key encryption
information on the expected upper and lower bound and a suggested rule based on these bounds. If generated, they are displayed with the Data Quality Results.		
Statistics Contains information on metrics collected after evaluating the data given a ruleset, such as the value of the metric (e.g. RowCount, Completeness), column names, and other metadata.	Enabled	Enabled
Anomaly Detection Statistic Models The statistic models contain the time series of upper and lower bounds for a given metric generated based off of previous evaluations of the customer data.	Enabled	Enabled

Note

AWS Data Quality automatically enables encryption at rest using AWS owned keys to protect personally identifiable data at no charge. However, AWS KMS charges apply for using a customer managed key. For more information about pricing, see AWS KMS Pricing. For more information on AWS KMS, see AWS KMS.

Create a Customer Managed Key

You can create a symmetric customer managed key by using the AWS Management Console, or the AWS KMS APIs.

To create a symmetric customer managed key:

- Follow the steps for [Creating symmetric encryption AWS KMS keys] in the AWS Key Management Service Developer Guide.

Key policy

Key policies control access to your customer managed key. Every customer managed key must have exactly one key policy, which contains statements that determine who can use the key and how they can use it. When you create your customer managed key, you can specify a key policy. For more information, see Key policies in AWS KMS keys in the AWS Key Management Service Developer Guide.

To use your customer managed key with your Data Quality resources, the following API operations must be permitted in the key policy:

- `kms:Decrypt` – Decrypts ciphertext that was encrypted by an AWS KMS key using `GenerateDataKeyWithoutPlaintext`
- `kms:DescribeKey` – Provides the customer managed key details to allow Amazon Location to validate the key.
- `kms:GenerateDataKeyWithoutPlaintext` – Returns a unique symmetric data key for use outside of AWS KMS. This operation returns a data key that is encrypted under a symmetric encryption KMS key that you specify. The bytes in the key are random; they are not related to the caller or to the KMS key. Used to reduce KMS calls the customer has to make.
- `kms:ReEncrypt*` – Decrypts ciphertext and then reencrypts it entirely within AWS KMS. You can use this operation to change the KMS key under which data is encrypted, such as when you manually rotate a KMS key or change the KMS key that protects a ciphertext. You can also use it to reencrypt ciphertext under the same KMS key, such as to change the encryption context of a ciphertext.

The following are policy statement examples you can add for Amazon Location:

```
"Statement" : [
    {
        "Sid" : "Allow access to principals authorized to use AWS Glue Data Quality",
        "Effect" : "Allow",
        "Principal" : {
            "AWS": "arn:aws:iam::<account_id>:role/ExampleRole"
        },
        "Action" : [
            "kms:Decrypt",
            "kms:DescribeKey",
            "kms:GenerateDataKeyWithoutPlaintext",
            "kms:ReEncrypt*"
        ],
        "Resource" : "*",
        "Condition" : {
            "StringEquals" : {
                "kms:ViaService" : "glue.amazonaws.com",
                "kms:CallerAccount" : "111122223333"
            }
        },
    {
        "Sid": "Allow access for key administrators",
        "Effect": "Allow",
        "Principal": {
            "AWS": "arn:aws:iam::111122223333:root"
          },
        "Action" : [
            "kms:*"
         ],
        "Resource": "arn:aws:kms:region:111122223333:key/key_ID"
    },
    {
        "Sid" : "Allow read-only access to key metadata to the account",
        "Effect" : "Allow",
        "Principal" : {
            "AWS" : "arn:aws:iam::111122223333:root"
        },
        "Action" : [
            "kms:Describe*",
            "kms:Get*",
            "kms:List*",
        ],
        "Resource" : "*"
    }
]
```

Notes about using KMS Keys in AWS Glue Data Quality

AWS Glue Data Quality does not support key transitions. This means that if you encrypt your Data Quality assets with Key A and decide to switch to Key B, we will not re-encrypt the data that was encrypted with Key A to use Key B. You can still switch to Key B, but will need to maintain access to Key A to access data previously encrypted with Key A.

For more information about specifying permissions in a policy, see [Permissions for AWS services in key policies] in the AWS Key Management Service Developer Guide.

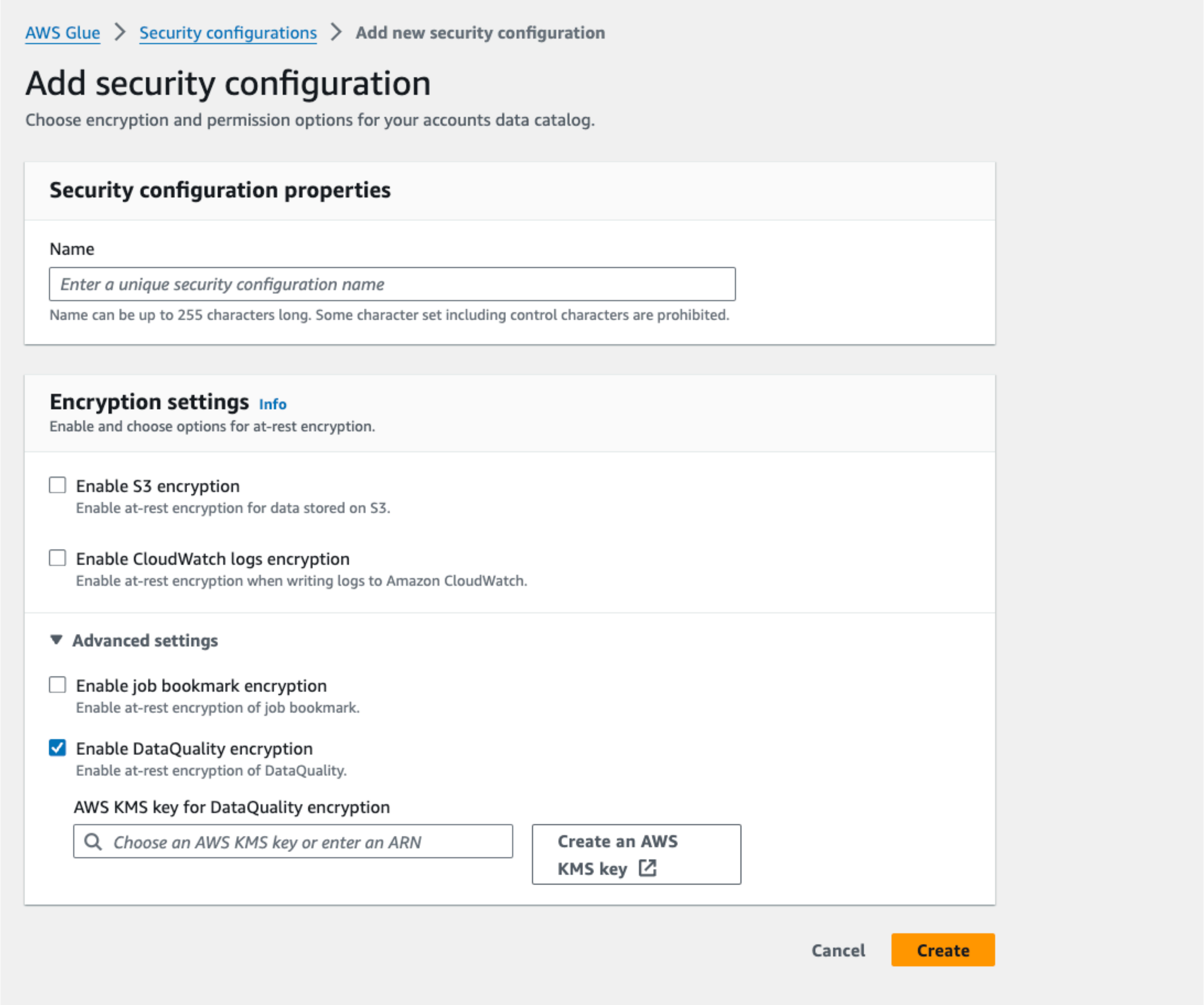
For more information about troubleshooting key access, see [Troubleshooting key access] in the AWS Key Management Service Developer Guide.

Create a security configuration

In AWS Glue the Security Configurations resource contains properties that are needed when you write encrypted data.

To encrypt your data quality assets:

1. In **Encryption settings**, under **Advanced settings**, choose **Enable Data Quality Encryption**
2. Select your KMS key or choose **Create an AWS KMS key**


AWS Glue > Security configurations > Add new security configuration

Add security configuration

Choose encryption and permission options for your accounts data catalog.

Security configuration properties

Name

Enter a unique security configuration name

Name can be up to 255 characters long. Some character set including control characters are prohibited.

Encryption settings Info

Enable and choose options for at-rest encryption.

☐ Enable S3 encryption
Enable at-rest encryption for data stored on S3.

☐ Enable CloudWatch logs encryption
Enable at-rest encryption when writing logs to Amazon CloudWatch.

▼ Advanced settings

☐ Enable job bookmark encryption
Enable at-rest encryption of job bookmark.

☑ Enable DataQuality encryption
Enable at-rest encryption of DataQuality.

AWS KMS key for DataQuality encryption

Choose an AWS KMS key or enter an ARN

AWS Glue Data Quality encryption context

An encryption context is an optional set of key-value pairs that contain additional contextual information about the data.

AWS KMS uses the encryption context as additional authenticated data to support authenticated encryption. When you include an encryption context in a request to encrypt data, AWS KMS binds the encryption context to the encrypted data. To decrypt data, you include the same encryption context in the request.

AWS Glue Data Quality encryption context example

```
"encryptionContext": {
    "kms-arn": "arn:aws:kms:us-
east-1:111122223333:key/1234abcd-12ab-34cd-56ef-123456SAMPLE",
    "branch-key-id": "111122223333+arn:aws:kms:us-
east-1:111122223333:key/1234abcd-12ab-34cd-56ef-123456SAMPLE",
    "hierarchy-version": "1",
    "aws-crypto-ec:aws:glue:securityConfiguration": "111122223333:customer-security-
configuration-name",
    "create-time": "2024-06-07T13:47:23:000861Z",
    "tablename": "AwsGlueMlEncryptionKeyStore",
    "type": "beacon:ACTIVE"
}
```

Using encryption context for monitoring

When you use a symmetric customer managed key to encrypt your tracker or geofence collection, you can also use the encryption context in audit records and logs to identify how the customer managed key is being used. The encryption context also appears in logs generated by AWS CloudTrail or Amazon CloudWatch Logs.

Monitoring your encryption keys for AWS Glue Data Quality

When you use an AWS KMS customer managed key with your AWS Glue Data Quality resources, you can use AWS CloudTrail or Amazon CloudWatch Logs to track requests that AWS Glue Data Quality sends to AWS KMS.

The following examples are AWS CloudTrail events for `GenerateDataKeyWithoutPlainText` and `Decrypt` to monitor KMS operations called by AWS Glue Data Quality to access data encrypted by your customer managed key.

Decrypt

```
{
    "eventVersion": "1.09",
    "userIdentity": {
        "type": "AssumedRole",
        "arn": "arn:aws:sts::111122223333:role/CustomerRole",
        "accountId": "111122223333",
        "invokedBy": "glue.amazonaws.com"
    },
    "eventTime": "2024-07-02T20:03:10Z",
    "eventSource": "kms.amazonaws.com",
    "eventName": "Decrypt",
    "awsRegion": "us-east-1",
    "sourceIPAddress": "glue.amazonaws.com",
    "userAgent": "glue.amazonaws.com",
    "requestParameters": {
        "keyId": "arn:aws:kms:us-
east-1:111122223333:key/1234abcd-12ab-34cd-56ef-123456SAMPLE",
        "encryptionAlgorithm": "SYMMETRIC_DEFAULT",
        "encryptionContext": {
            "kms-arn": "arn:aws:kms:us-
east-1:111122223333:key/1234abcd-12ab-34cd-56ef-123456SAMPLE",
            "branch-key-id": "111122223333+arn:aws:kms:us-
east-1:111122223333:key/1234abcd-12ab-34cd-56ef-123456SAMPLE",
            "hierarchy-version": "1",
            "aws-crypto-ec:aws:glue:securityConfiguration": "111122223333:customer-
security-configuration-name",
            "create-time": "2024-06-07T13:47:23:000861Z",
            "tablename": "AwsGlueMlEncryptionKeyStore",
            "type": "branch:ACTIVE",
            "version": "branch:version:ff000af-00eb-00ce-0e00-ea000fb0fba0SAMPLE"
        }
    },
    "responseElements": null,
    "requestID": "ff000af-00eb-00ce-0e00-ea000fb0fba0SAMPLE",
    "eventID": "ff000af-00eb-00ce-0e00-ea000fb0fba0SAMPLE",
    "readOnly": true,
    "resources": [
        {
            "accountId": "111122223333",
            "type": "AWS::KMS::Key",
            "ARN": "arn:aws:kms:us-
east-1:111122223333:key/1234abcd-12ab-34cd-56ef-123456SAMPLE"
        }
    ],
    "eventType": "AwsApiCall",
    "managementEvent": true,
    "recipientAccountId": "111122223333",
    "eventCategory": "Management"
}
```

GenerateDataKeyWithoutPlaintext

```
{
    "eventVersion": "1.09",
    "userIdentity": {
        "type": "AssumedRole",
        "arn": "arn:aws:sts::111122223333:role/CustomerRole",
        "accountId": "111122223333",
        "invokedBy": "glue.amazonaws.com"
    },
    "eventTime": "2024-07-02T20:03:10Z",
    "eventSource": "kms.amazonaws.com",
    "eventName": "GenerateDataKeyWithoutPlaintext",
    "awsRegion": "us-east-1",
    "sourceIPAddress": "glue.amazonaws.com",
    "userAgent": "glue.amazonaws.com",
    "requestParameters": {
        "keyId": "arn:aws:kms:us-
east-1:111122223333:key/1234abcd-12ab-34cd-56ef-123456SAMPLE",
        "encryptionAlgorithm": "SYMMETRIC_DEFAULT",
        "encryptionContext": {
            "kms-arn": "arn:aws:kms:us-
east-1:111122223333:key/1234abcd-12ab-34cd-56ef-123456SAMPLE",
            "branch-key-id": "111122223333+arn:aws:kms:us-
east-1:111122223333:key/1234abcd-12ab-34cd-56ef-123456SAMPLE",
            "hierarchy-version": "1",
            "aws-crypto-ec:aws:glue:securityConfiguration": "111122223333:customer-
security-configuration-name",
            "create-time": "2024-06-07T13:47:23:000861Z",
            "tablename": "AwsGlueMlEncryptionKeyStore",
            "type": "branch:version:ff000af-00eb-00ce-0e00-ea000fb0fba0SAMPLE"
        }
    },
    "responseElements": null,
    "requestID": "ff000af-00eb-00ce-0e00-ea000fb0fba0SAMPLE",
    "eventID": "ff000af-00eb-00ce-0e00-ea000fb0fba0SAMPLE",
    "readOnly": true,
    "resources": [
        {
            "accountId": "111122223333",
            "type": "AWS::KMS::Key",
            "ARN": "arn:aws:kms:us-
east-1:111122223333:key/1234abcd-12ab-34cd-56ef-123456SAMPLE"
        }
    ],
    "eventType": "AwsApiCall",
    "managementEvent": true,
    "recipientAccountId": "111122223333",
    "eventCategory": "Management"
}
```

ReEncyrpt

```
{
    "eventVersion": "1.09",
    "userIdentity": {
        "type": "AssumedRole",
        "arn": "arn:aws:sts::111122223333:role/CustomerRole",
        "accountId": "111122223333",
        "invokedBy": "glue.amazonaws.com"
    },
    "eventTime": "2024-07-17T21:34:41Z",
    "eventSource": "kms.amazonaws.com",
    "eventName": "ReEncrypt",
    "awsRegion": "us-east-1",
    "sourceIPAddress": "glue.amazonaws.com",
    "userAgent": "glue.amazonaws.com",
    "requestParameters": {
        "destinationEncryptionContext": {
             "kms-arn": "arn:aws:kms:us-
east-1:111122223333:key/1234abcd-12ab-34cd-56ef-123456SAMPLE",
            "branch-key-id": "111122223333+arn:aws:kms:us-
east-1:111122223333:key/1234abcd-12ab-34cd-56ef-123456SAMPLE",
            "hierarchy-version": "1",
            "aws-crypto-ec:aws:glue:securityConfiguration": "111122223333:customer-
security-configuration-name",
            "create-time": "2024-06-07T13:47:23:000861Z",
            "tablename": "AwsGlueMlEncryptionKeyStore",
            "type": "branch:ACTIVE"
            "version": "branch:version:12345678-SAMPLE"
        },
        "destinationKeyId": "arn:aws:kms:us-
east-1:111122223333:key/1234abcd-12ab-34cd-56ef-123456SAMPLE",
        "sourceAAD": "1234567890-SAMPLE+Z+lqoYOHj7VtWxJLrvh+biUFbliYDAQkobM=",
        "sourceKeyId": "arn:aws:kms:ap-southeast-2:585824196334:key/17ca05ca-a8c1-40d7-
b7fd-30abb569a53a",
        "destinationEncryptionAlgorithm": "SYMMETRIC_DEFAULT",
        "sourceEncryptionContext": {
            "kms-arn": "arn:aws:kms:us-
east-1:111122223333:key/1234abcd-12ab-34cd-56ef-123456SAMPLE",
            "branch-key-id": "111122223333+arn:aws:kms:us-
east-1:111122223333:key/1234abcd-12ab-34cd-56ef-123456SAMPLE",
            "hierarchy-version": "1",
            "aws-crypto-ec:aws:glue:securityConfiguration": "111122223333:customer-
security-configuration-name",
            "create-time": "2024-06-07T13:47:23:000861Z",
            "tablename": "AwsGlueMlEncryptionKeyStore",
            "type": "branch:version:ff000af-00eb-00ce-0e00-ea000fb0fba0SAMPLE"
        },
        "destinationAAD": "1234567890-SAMPLE",
        "sourceEncryptionAlgorithm": "SYMMETRIC_DEFAULT"
    },
    "responseElements": null,
    "requestID": "ff000af-00eb-00ce-0e00-ea000fb0fba0SAMPLE",
    "eventID": "ff000af-00eb-00ce-0e00-ea000fb0fba0SAMPLE",
    "readOnly": true,
    "resources": [
        {
            "accountId": "111122223333",
            "type": "AWS::KMS::Key",
            "ARN": "arn:aws:kms:us-
east-1:111122223333:key/1234abcd-12ab-34cd-56ef-123456SAMPLE"
        },
        {
            "accountId": "111122223333",
            "type": "AWS::KMS::Key",
            "ARN": "arn:aws:kms:us-
east-1:111122223333:key/1234abcd-12ab-34cd-56ef-123456SAMPLE"
        }
    ],
    "eventType": "AwsApiCall",
    "managementEvent": true,
    "recipientAccountId": "111122223333",
    "eventCategory": "Management"
}
```

Learn more

The following resources provide more information about data encryption at rest.

- For more information about AWS Key Management Service basic concepts see the AWS Key Management Service Developer Guide.
- For more information about Security best practices for AWS Key Management Service see the AWS Key Management Service Developer Guide.

Troubleshooting AWS Glue Data Quality errors

If you encounter errors in AWS Glue Data Quality, use the following solutions to help you find the source of the problems and fix them.

Contents

- Error: missing AWS Glue Data Quality module
- Error: insufficient AWS Lake Formation permissions
- Error: rulesets are not uniquely named
- Error: tables with special characters
- Error: overflow error with a large ruleset
- Error: overall rule status is failed
- AnalysisException: Unable to verify existence of default database
- Error Message: Provided key map not suitable for given data frames
- Exception in User Class: java.lang.RuntimeException : Failed to fetch data. Check the logs in CloudWatch to get more details
- LAUNCH ERROR: Error downloading from S3 for bucket
- InvalidInputException (status: 400): DataQuality rules cannot be parsed
- Error: Eventbridge is not triggering Glue DQ jobs based on the schedule I setup
- CustomSQL errors
- Dynamic Rules
- Exception in User Class: org.apache.spark.sql.AnalysisException: org.apache.hadoop.hive.ql.metadata.HiveException
- UNCLASSIFIED_ERROR; IllegalArgumentException: Parsing Error: No rules or analyzers provided., no viable alternative at input

Error: missing AWS Glue Data Quality module

Error message: No module named 'awsgluedq'.

Resolution: This error occurs when you run AWS Glue Data Quality in an unsupported version. AWS Glue Data Quality is supported only in Glue version 3.0 and later.

Error: insufficient AWS Lake Formation permissions

Error message: Exception in User Class: `com.amazonaws.services.glue.model.AccessDeniedException`: Insufficient Lake Formation permission(s) on impact_sdg_involvement (Service: AWS Glue; Status Code: 400; Error Code: AccessDeniedException; Request ID: 465ae693-b7ba-4df0-a4e4-6b17xxxxxxxx; Proxy: null).

Resolution: You must provide sufficient permissions in AWS Lake Formation.

Error: rulesets are not uniquely named

Error message: Exception in User Class: ...services.glue.model.AlreadyExistsException: Another ruleset with the same name already exists.

Resolution: Rulesets are global and must be unique.

Error: tables with special characters

Error message: Exception in User Class: org.apache.spark.sql.AnalysisException: cannot resolve ''C'' given input columns: [primary.data_end_time, primary.data_start_time, primary.end_time, primary.last_updated, primary.message, primary.process_date, primary.rowhash, primary.run_by, primary.run_id, primary.start_time, primary.status]; line 1 pos 44;.

Resolution: There is a current limitation that AWS Glue Data Quality cannot be executed on tables that have special characters such as ".".

Error: overflow error with a large ruleset

Error message: Exception in User Class: java.lang.StackOverflowError.

Resolution: If you have a large ruleset of greater than 2K rules, you may encounter this issue. Break your rules into multiple rulesets.

Error: overall rule status is failed

Error condition: My Ruleset is successful, but my overall rule status is failed.

Resolution: This error most likely occurred because you chose the option to publish metrics to Amazon CloudWatch while publishing. If your dataset is in a VPC, your VPC may not allow AWS Glue to publish metrics to Amazon CloudWatch. In this case, you >must set up an endpoint for your VPC to access Amazon CloudWatch.

AnalysisException: Unable to verify existence of default database

Error condition: AnalysisException: Unable to verify existence of default database: com.amazonaws.services.glue.model.AccessDeniedException: Insufficient Lake Formation permission(s) on default (Service: AWS Glue; Status Code: 400; Error Code: AccessDeniedException; Request ID: XXXXXXXX-XXXX-XXXX-XXXX-XXXXXXXXXXXX; Proxy: null)

Resolution: In AWS Glue job's catalog integration, AWS Glue always tries to check whether default database exists or not using AWS Glue `GetDatabase API`. When the `DESCRIBE` Lake Formation permission is not granted, or `GetDatabase IAM` permission is granted, then the job fails when verifying existence of the default database.

To resolve:

1. Add the `DESCRIBE` permission in Lake Formation for the default database.
2. Configure the IAM role attached to the AWS Glue job as Database Creator in Lake Formation. This will automatically create a default database and grant required Lake Formation permissions for the role.
3. Disable `--enable-data-catalog` option. (It is shown as **Use Data Catalog as the Hive metastore** in AWS Glue Studio).

   If you do not need Spark SQL Data Catalog integration in the job, you can disable it.

Error Message: Provided key map not suitable for given data frames

Error condition: Provided key map not suitable for given data frames.

Resolution: You are using **DataSetMatch** ruletype and the join keys have duplicates. Your join keys must be unique and must not be NULL. In cases where you can't have join keys that are unique, consider using other ruletypes such as **AggregateMatch** to match on summary data.

Exception in User Class: java.lang.RuntimeException : Failed to fetch data. Check the logs in CloudWatch to get more details

Error condition: Exception in User Class: java.lang.RuntimeException : Failed to fetch data. Check the logs in CloudWatch to get more details.

Resolution: This happens when you are creating DQ rules on an Amazon S3-based table that compares against Amazon RDS or Amazon Redshift. In these cases, AWS Glue cannot load the connection. Instead, try to set up DQ rule on the Amazon Redshift or Amazon RDS dataset. This is a known bug.

LAUNCH ERROR: Error downloading from S3 for bucket

Error condition: LAUNCH ERROR: Error downloading from S3 for bucket: `aws-glue-ml-data-quality-assets-us-east-1, key: jars/aws-glue-ml-data-quality-etl.jar.Access Denied (Service: Amazon S3; Status Code: 403; Please refer logs for details) .`

Resolution: The permissions in the role passed to AWS Glue Data Quality must permit reading from the preceding Amazon S3 location. This IAM policy should be attached to the role:

```
{
  "Sid": "allowS3",
  "Effect": "Allow",
  "Action": "s3:GetObject",
  "Resource": "arn:aws:s3:::aws-glue-ml-data-quality-assets-<region>/*"
}
```

Refer to Data Quality authorization for detailed permissions. These libraries are required to evaluate data quality for your datasets.

InvalidInputException (status: 400): DataQuality rules cannot be parsed

Error condition: InvalidInputException (status: 400): DataQuality rules cannot be parsed.

Resolution: There are many possibilities for this error. One possibility is that your rules may have single quotes. Verify that they are in double quotes. For example:

```
Rules = [
ColumnValues "tipo_vinculo" in ["CODO", "DOCO", "COCO", "DODO"] AND "categoria" = 'ES"
    AND "cod_bandera" = 'CEP'
```

Change this to:

```
Rules = [
(ColumnValues "tipovinculo" in [ "CODO", "DOCO", "COCO", "DODO"]) AND (ColumnValues
 "categoria" = "ES")
    AND (ColumnValues "codbandera" = "CEP")
]
```

Error: Eventbridge is not triggering Glue DQ jobs based on the schedule I setup

Error condition: Eventbridge is not triggering AWS Glue Data Quality jobs based on the schedule I setup.

Resolution: The role triggering the job may not have the right permissions. Make sure that the role that you are using to start the jobs has the permissions mentioned in IAM setup required for scheduling evaluation runs .

CustomSQL errors

Error condition: `The output from CustomSQL must contain at least one column that matches the input dataset for AWS Glue Data Quality to provide row level results. The SQL query is a valid query but no columns from the SQL result are present in the Input Dataset. Ensure that matching columns are returned from the SQL.`

Resolution: The SQL query is valid but verify that you're selecting only columns from the primary table. Selecting aggregate functions like sum, count on the columns from the primary can result in this error.

Error condition: `There was a problem when executing your SQL statement: cannot resolve "Col".`

Resolution: This column is not present in the primary table.

Error: Eventbridge is not triggering Glue DQ jobs based on the schedule I setup 2048

Error condition: `The columns that are returned from the SQL statement should only belong to the primary table. "In this case, some columns ( Col ) belong to reference table"`.

Resolution: In SQL queries when you're joining the primary table with other reference tables, verify that your select statement has only column names from your primary table to generate row level results for the primary table.

Dynamic Rules

Error condition: `Dynamic rules require job context, and cannot be evaluated in interactive session or data preview..`

Cause: This error message might appear in your data preview results, or in other interactive sessions, when dynamic DQ rules are present in your ruleset. Dynamic rules reference historical metrics associated with a particular job name and evaluation context, so they can't be evaluated in interactive sessions.

Resolution: Running your AWS Glue job will produce historical metrics, which can be referenced in later job runs for the same job.

Error condition:

- `[RuleType] rule only supports simple atomic operands in thresholds..`
- `Function last not yet implemented for [RuleType] rule.`

Resolution: Dynamic rules are generally supported for all DQDL ruletypes in numeric expressions (see DQDL Reference). However, some rules that produce multiple metrics, ColumnValues and ColumnLength, are not yet supported.

Error condition: `Binary expression operands must resolve to a single number..`

Cause: Dynamic rules support binary expressions, like `RowCount > avg(last(5)) * 0.9`. Here, the binary expression is `avg(last(5)) * 0.9`. This rule is valid because both operands `avg(last(5))` and `0.9` resolve to a single number. An incorrect example is `RowCount > last(5) * 0.9`, because `last(5)` will produce a list that can't be meaningfully compared to the current row count.

Resolution: Use aggregation functions to reduce a list-valued operand to a single number.

Error condition:

- `Rule threshold results in list, and a single value is expected. Use aggregation functions to produce a single value. Valid example: sum(last(10)), avg(last(10)).`
- `Rule threshold results in empty list, and a single value is expected.`

Cause: Dynamic rules can be used to compare some feature of your dataset with its historical values. The last function allows for the retrieval of multiple historical values, if a positive integer argument is provided. For example, `last(5)` will retrieve the last five most recent values observed in job runs for your rule.

Resolution: An aggregation function must be used to reduce these values to a single number to make a meaningful comparison with the value observed in the current job run.

Valid examples:

- `RowCount >= avg(last(5))`
- `RowCount > last(1)`
- `RowCount < last()`

Invalid example: `RowCount > last(5)`.

Error condition:

- `Function index used in threshold requires positive integer argument.`
- `Index argument must be an integer. Valid syntax example: RowCount > index(last(10, 2)), which means RowCount must be greater than third most recent execution from last 10 job runs.`

Resolution: When authoring dynamic rules, you can use the `index` aggregation function to select one historical value from a list. For example, `RowCount > index(last(5), 1)` will check whether the row count observed in the current job is strictly greater than the second most recent row count observed for your job. `index` is zero-indexed.

Error condition: `IllegalArgumentException: Parsing Error: Rule Type: DetectAnomalies is not valid`.

Resolution: Anomaly detection is only available in AWS Glue 4.0.

Error condition: `IllegalArgumentException: Parsing Error: Unexpected condition for rule of type ... no viable alternative at input ....`

Note: `...` is dynamic. Example: `IllegalArgumentException: Parsing Error: Unexpected condition for rule of type RowCount with number return type, line 4:19 no viable alternative at input '>last'.`

Resolution: Anomaly detection is only available in AWS Glue 4.0.

Exception in User Class: org.apache.spark.sql.AnalysisException: org.apache.hadoop.hive.ql.metadata.HiveException

Error condition: `Exception in User Class: org.apache.spark.sql.AnalysisException: org.apache.hadoop.hive.ql.metadata.HiveException: Unable to fetch table mailpiece_submitted. StorageDescriptor#InputFormat cannot be null for table: mailpiece_submitted (Service: null; Status Code: 0; Error Code: null; Request ID: null; Proxy: null)`

Cause: You are using Apache Iceberg in AWS Glue Data Catalog and the Input Format attribute in AWS Glue Data Catalog is empty.

Resolution: This issue occurs when you are using CustomSQL ruletype in your DQ rule. One way to fix this is to use “primary“ or add catalog name `glue_catalog.` to `<database>.<table> in Custom ruletype.`

UNCLASSIFIED_ERROR; IllegalArgumentException: Parsing Error: No rules or analyzers provided., no viable alternative at input

Error condition: `UNCLASSIFIED_ERROR; IllegalArgumentException: Parsing Error: No rules or analyzers provided., no viable alternative at input`

Resolution: DQDL is not parsable. There are a few instances where this can occur. If you are using composite rules, makes sure they have right parenthesis.

```
(RowCount >= avg(last(10)) * 0.6) and (RowCount <= avg(last(10)) * 1.4) instead of
      RowCount >= avg(last(10)) * 0.6 and RowCount <= avg(last(10)) * 1.4
```

Amazon Q data integration in AWS Glue

Amazon Q data integration in AWS Glue is a new generative AI capability of AWS Glue that enables data engineers and ETL developers to build data integration jobs using natural language. Engineers and developers can ask Amazon Q to author jobs, troubleshoot issues, and answer questions about AWS Glue and data integration.

What is Amazon Q?

> **Note**
>
> Powered by Amazon Bedrock: AWS implements automated abuse detection. Because Amazon Q data integration is built on Amazon Bedrock, users can take full advantage of the controls implemented in Amazon Bedrock to enforce safety, security, and the responsible use of artificial intelligence (AI).

Amazon Q is a generative artificial intelligence (AI) powered conversational assistant that can help you understand, build, extend, and operate AWS applications. The model that powers Amazon Q has been augmented with high quality AWS content to get you more complete, actionable, and referenced answers to accelerate your building on AWS. For more information, see What is Amazon Q?

What is Amazon Q data integration in AWS Glue?

Amazon Q data integration in AWS Glue includes the following capabilities:

- **Chat** – Amazon Q data integration in AWS Glue can answer natural language questions in English about AWS Glue and data integration domains like AWS Glue source and destination connectors, AWS Glue ETL jobs, Data Catalog, crawlers and AWS Lake Formation, and other feature documentation, and best practices. Amazon Q data integration in AWS Glue responds with step-by-step instructions, and includes references to its information sources.
- **Data integration code generation** – Amazon Q data integration in AWS Glue can answer questions about AWS Glue ETL scripts, and generate new code given a natural language question in English.

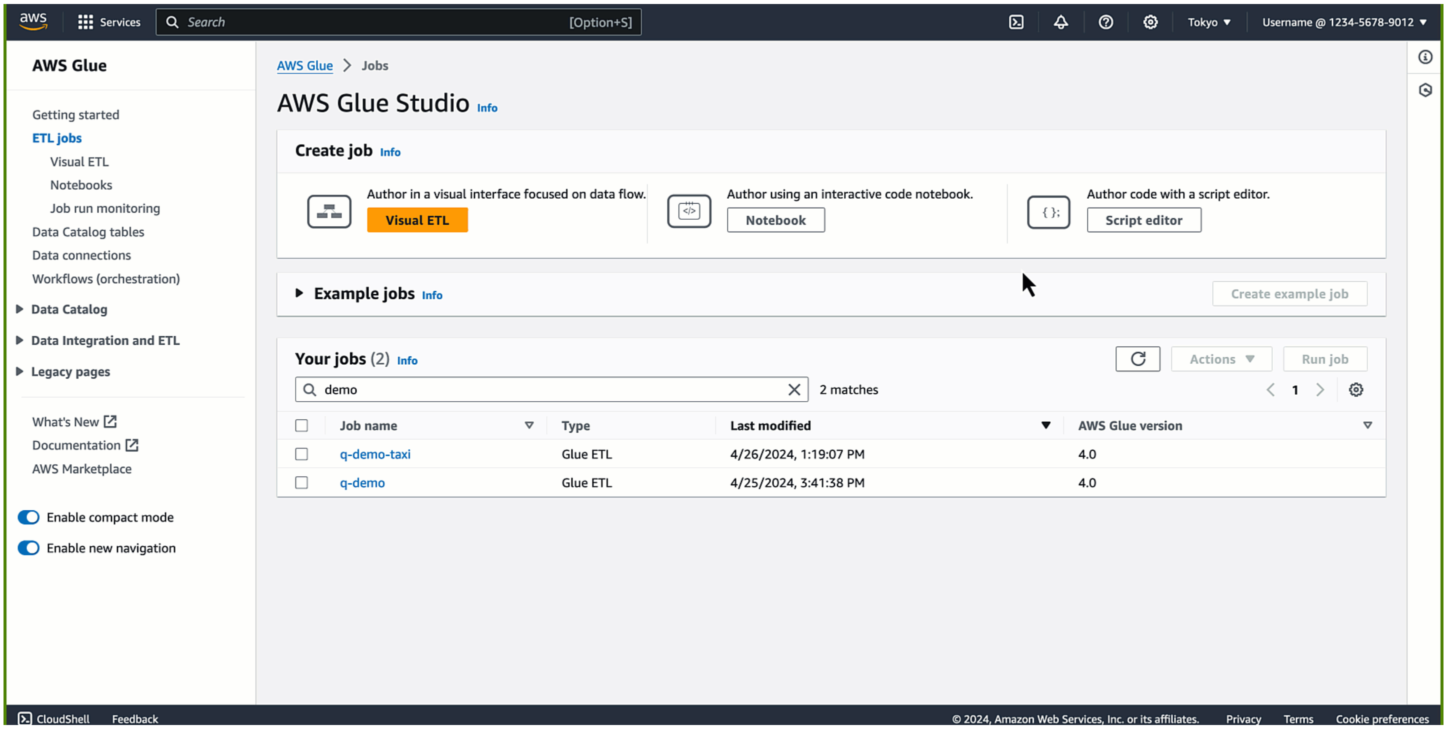
- **Troubleshoot** – Amazon Q data integration in AWS Glue is purpose built to help you understand errors in AWS Glue jobs and provides step-by-step instructions, to root cause and resolve your issues.

> **Note**
>
> Amazon Q data integration in AWS Glue does not use the context of your conversation to inform future responses for the duration of your conversation. Each conversation with Amazon Q data integration in AWS Glue is independent of your prior or future conversations.

Working with Amazon Q data integration in AWS Glue?

In the Amazon Q panel you can request Amazon Q generate code for an AWS Glue ETL script, or answer a question on AWS Glue features or troubleshooting an error. The response is an ETL script in PySpark with step-by-step instructions to customize the script, review and execute it. For questions, the response is generated based on the data integration knowledge base with a summary and source URL for references.

For example, you can ask Amazon Q to "*Please provide a Glue script that reads from Snowflake, renames the fields, and writes to Redshift*" and in response, Amazon Q data integration in AWS Glue will return an AWS Glue job script that can perform the requested action. You can review the generated code to ensure that it fulfills the requested intent. If satisfied, you can deploy it as an AWS Glue job in production. You can troubleshoot jobs by asking the integration to explain errors and failures, and to propose solutions. Amazon Q can answer questions about AWS Glue or data integration best practices.


Services
Search [Option+S]
Tokyo
Username @ 1234-5678-9012
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Data Catalog
Data Integration and ETL
Legacy pages
What's New
Documentation
AWS Marketplace
Enable compact mode
Enable new navigation
AWS Glue > Jobs
AWS Glue Studio Info
Create job Info
Author in a visual interface focused on data flow.
Visual ETL
Author using an interactive code notebook.
Notebook
Author code with a script editor.
Script editor
Example jobs Info
Create example job
Your jobs (2) Info
Actions
Run job
demo
2 matches
1
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

☐	Job name	Type	Last modified	AWS Glue version
☐	q-demo-taxi	Glue ETL	4/26/2024, 1:19:07 PM	4.0
☐	q-demo	Glue ETL	4/25/2024, 3:41:38 PM	4.0

The following are example questions that demonstrate how Amazon Q data integration in AWS Glue can help you build on AWS Glue:

AWS Glue ETL code generation:

- Write an AWS Glue script that reads JSON from S3, transforms fields using apply mapping and writes to Amazon Redshift
- How do I write an AWS Glue script for reading from DynamoDB, applying the DropNullFields transform and writing to S3 as Parquet?
- Give me an AWS Glue script that reads from MySQL, drops some fields based on my business logic, and writes to Snowflake
- Write an AWS Glue job to read from DynamoDB and write to S3 as JSON
- Help me develop an AWS Glue script for AWS Glue Data Catalog to S3
- Write an AWS Glue job to read JSON from S3, drop nulls and write to Redshift

AWS Glue feature explanations:

- How do I use AWS Glue Data Quality?
- How to use AWS Glue job bookmarks?
- How do I enable AWS Glue autoscaling?

- What is the difference between AWS Glue dynamic frames and Spark data frames?
- What are the different types of connections supported by AWS Glue?

AWS Glue troubleshooting:

- How to troubleshoot Out Of Memory (OOM) errors on AWS Glue jobs?
- What are some error messages you may see when setting up AWS Glue Data Quality and how can you fix them?
- How do I fix an AWS Glue job with the error Amazon S3 access denied?
- How do I resolve issues with data shuffle on AWS Glue jobs?

Best practices for interacting with Amazon Q data integration

The following are best practices for interacting with Amazon Q data integration:

- When interacting with Amazon Q data integration, ask specific questions, iterate when you have complex requests, and verify the answers for accuracy.
- When providing data integration prompts in natural language, be as specific as possible to help the assistant understand exactly what you need. Instead of asking "extract data from S3," provide more details like “write an AWS Glue script that extracts JSON files from S3.”
- Review the generated script before running it to ensure accuracy. If the generated script has errors or does not match your intent, provide instructions to the assistant on how to correct it.
- Generative AI technology is new and there can be mistakes, sometimes called hallucinations, in the responses. Test and review all code for errors and vulnerabilities before using it in your environment or workload.

Amazon Q data integration in AWS Glue service improvement

To help Amazon Q data integration in AWS Glue provide the most relevant information about AWS services, we may use certain content from Amazon Q, such as questions that you ask Amazon Q and its responses, for service improvement.

For information about what content we use and how to opt out, see Amazon Q Developer service improvement in the *Amazon Q Developer User Guide*.

Considerations

Consider the following items before you use Amazon Q data integration in AWS Glue:

- Currently, the code generation only works with PySpark kernel. The generated code is for AWS Glue jobs based on Python Spark.
- For information about the supported combinations of code generation abilities of Amazon Q data integration in AWS Glue, see Supported code generation abilities.

Setting up Amazon Q data integration in AWS Glue

The following sections provide information setting up Amazon Q data integration in AWS Glue.

Topics

- Configuring IAM permissions

Configuring IAM permissions

This topic describes the IAM permissions that you configure for the Amazon Q chat experience, and the AWS Glue Studio notebook experience.

Topics

- Configuring IAM permissions for Amazon Q chat
- Configuring IAM permissions for AWS Glue Studio notebooks

Configuring IAM permissions for Amazon Q chat

Granting permissions to the APIs used by Amazon Q data integration in AWS Glue requires appropriate AWS Identity and Access Management (IAM) permissions. You can obtain permissions by attaching the following custom AWS policy to your IAM identity (such as a user, role, or group):

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:StartCompletion",
        "glue:GetCompletion"
      ],
      "Resource": [
        "arn:aws:glue:*:*:completion/*"
      ]
    }
  ]
}
```

Configuring IAM permissions for AWS Glue Studio notebooks

To enable Amazon Q data integration in AWS Glue Studio notebooks, ensure the following permission is attached to the notebook IAM role:

> **Note**
>
> The `codewhisperer` prefix is a legacy name from a service that merged with Amazon Q Developer. For more information, see [Amazon Q Developer rename - Summary of changes](#).

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:StartCompletion",
        "glue:GetCompletion"
      ],
      "Resource": [
        "arn:aws:glue:*:*:completion/*"
      ]
    },
    {
      "Sid": "AmazonQDeveloperPermissions",
      "Effect": "Allow",
      "Action": [
        "codewhisperer:GenerateRecommendations"
      ],
      "Resource": "*"
    }
  ]
}
```

> **Note**
>
> Amazon Q data integration in AWS Glue does not have APIs available through the AWS SDK that you can use programmatically. The following two APIs are used in the IAM policy for enabling this experience through the Amazon Q chat panel or AWS Glue Studio notebooks: `StartCompletion` and `GetCompletion`.

Assigning permissions

To provide access, add permissions to your users, groups, or roles:

- Users and groups in AWS IAM Identity Center: Create a permission set. Follow the instructions in Create a permission set in the *AWS IAM Identity Center User Guide*.
- Users managed in IAM through an identity provider: Create a role for identity federation. Follow the instructions in Creating a role for a third-party identity provider (federation) in the *IAM User Guide*.
- IAM users:
  - Create a role that your user can assume. Follow the instructions in Creating a role for an IAM user in the *IAM User Guide*.
  - (Not recommended) Attach a policy directly to a user or add a user to a user group. Follow the instructions in Adding permissions to a user (console) in the *IAM User Guide*.

Supported code generation abilities

The following are the combinations of the code generation abilities of Amazon Q data integration.

Sources and Targets	Transformation
S3 with the following format types: json, csv, parquet, hudi, delta	Drop
AWS Glue Data Catalog	Aggregate
Redlake	DropDuplicates
Amazon DynamoDB	Join
MySQL	Filter
Oracle	RenameColumns
PostgresSQL	FillNull
Microsoft SQL Server	DropNull
Amazon DocumentDB / MongoDB	WithColumns
Snowflake	SQL Query
Google BigQuery	Union
Teradata	Select
Amazon OpenSearch Service	
Vertica	
SAP HANA	
Amazon Redshift	

Example interactions

Amazon Q data integration in AWS Glue allows you enter your question in the Amazon Q panel. You can enter a question regarding data integration functionality provided by AWS Glue. A detailed answer, together with reference documents, will be returned.

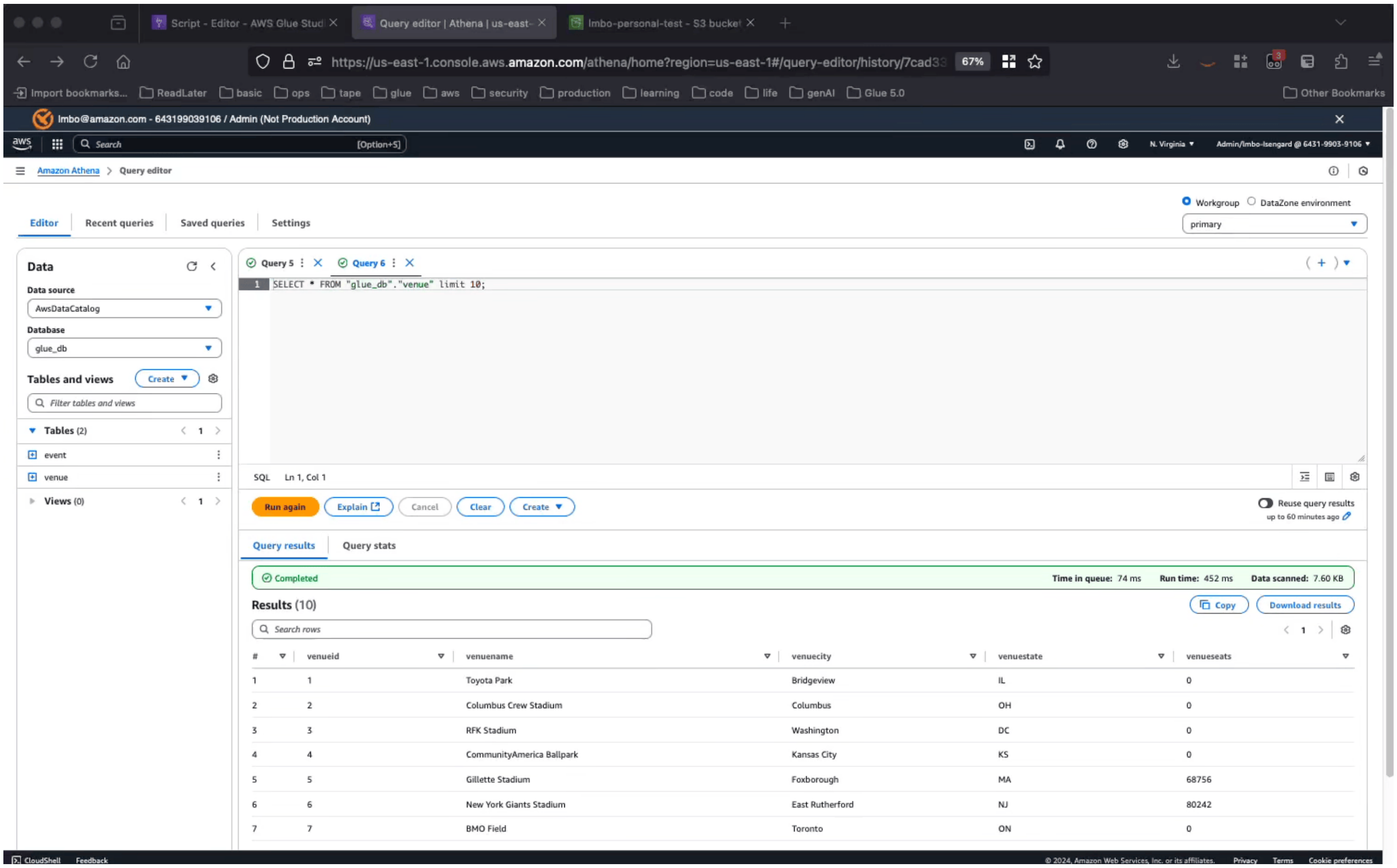
Another use case is generating AWS Glue ETL job scripts. You can ask a question regarding how to perform a data extract, transform, load job. A generated PySpark script will be returned.

Topics

- Amazon Q chat interactions
- AWS Glue Studio notebook interactions

Amazon Q chat interactions

On the AWS Glue console, start authoring a new job, and ask Amazon Q: "*Create a Glue ETL flow connect to two Glue catalog tables venue and event in my database glue_db, join the results on the venue's venueid and event's e_venueid, and then filter on venue state with condition as venuestate=='DC' and write to s3://amzn-s3-demo-bucket/codegen/BDB-9999/output/ in CSV format.*""


Amazon Athena > Query editor
Editor
Recent queries
Saved queries
Settings
Workgroup
DataZone environment
primary
Data
Data source
AwsDataCatalog
Database
glue_db
Tables and views
Create
Tables (2)
event
venue
Views (0)
Query 5
Query 6
Run again
Explain
Cancel
Clear
Create
Reuse query results up to 60 minutes ago
Query results
Query stats
Copy
Download results

```
SELECT * FROM "glue_db"."venue" limit 10;
```

Completed Time in queue: 74 ms Run time: 452 ms Data scanned: 7.60 KB

Results (10)

#	venueid	venuename	venuecity	venuestate	venueseats
1	1	Toyota Park	Bridgeview	IL	0
2	2	Columbus Crew Stadium	Columbus	OH	0
3	3	RFK Stadium	Washington	DC	0
4	4	CommunityAmerica Ballpark	Kansas City	KS	0
5	5	Gillette Stadium	Foxborough	MA	68756
6	6	New York Giants Stadium	East Rutherford	NJ	80242
7	7	BMO Field	Toronto	ON	0

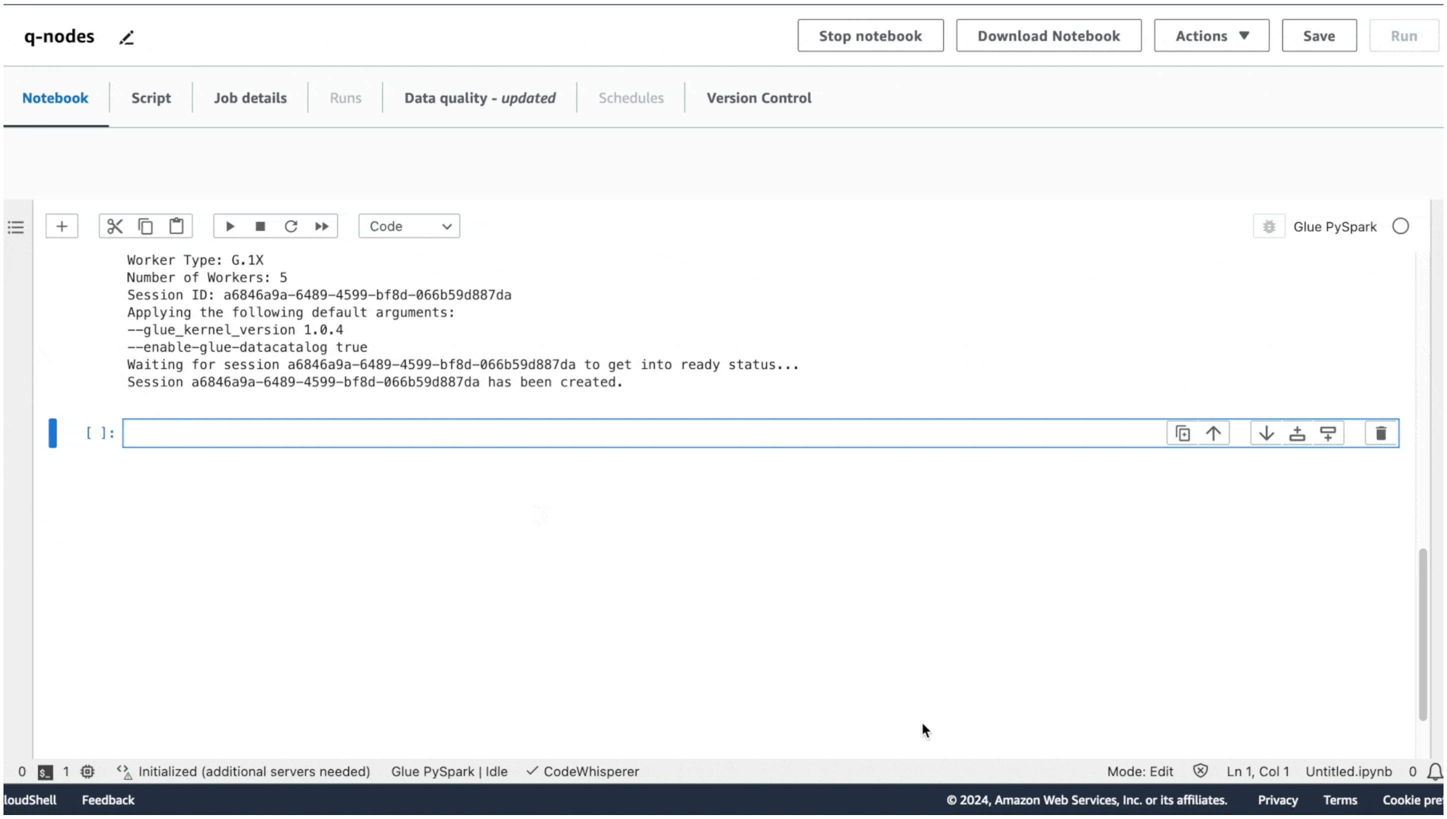
You will notice that the code is generated. With this response, you can learn and understand how you can author AWS Glue code for your purpose. You can copy/paste the generated code to the script editor and configure placeholders. After you configure an IAM role and AWS Glue connections on the job, save and run the job. When the job is complete, you can verify the summary data is persisted to Amazon S3 as expected and can be used by your downstream workloads.

AWS Glue Studio notebook interactions

> **Note**
>
> The Amazon Q Data integration experience in AWS Glue Studio notebook still focuses on DynamicFrame-based data integration flow.

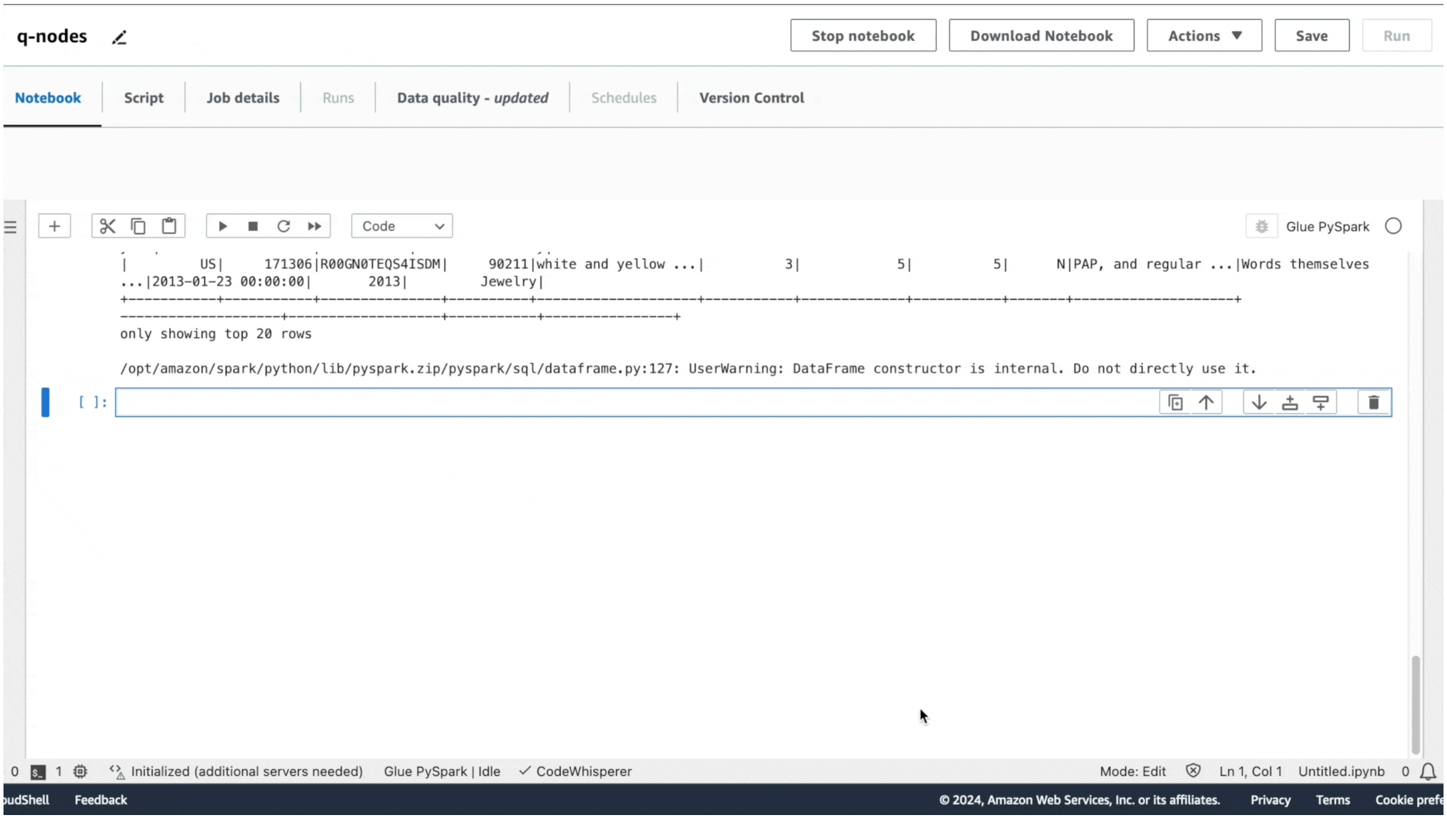
Add a new cell and enter your comment to describe what you want to achieve. After you press **Tab** and **Enter**, the recommended code is shown.

First intent is to extract the data: *"Give me code that reads a Glue Data Catalog table"*, followed by *"Give me code to apply a filter transform with star_rating>3"* and *"Give me code that writes the frame into S3 as Parquet"*.

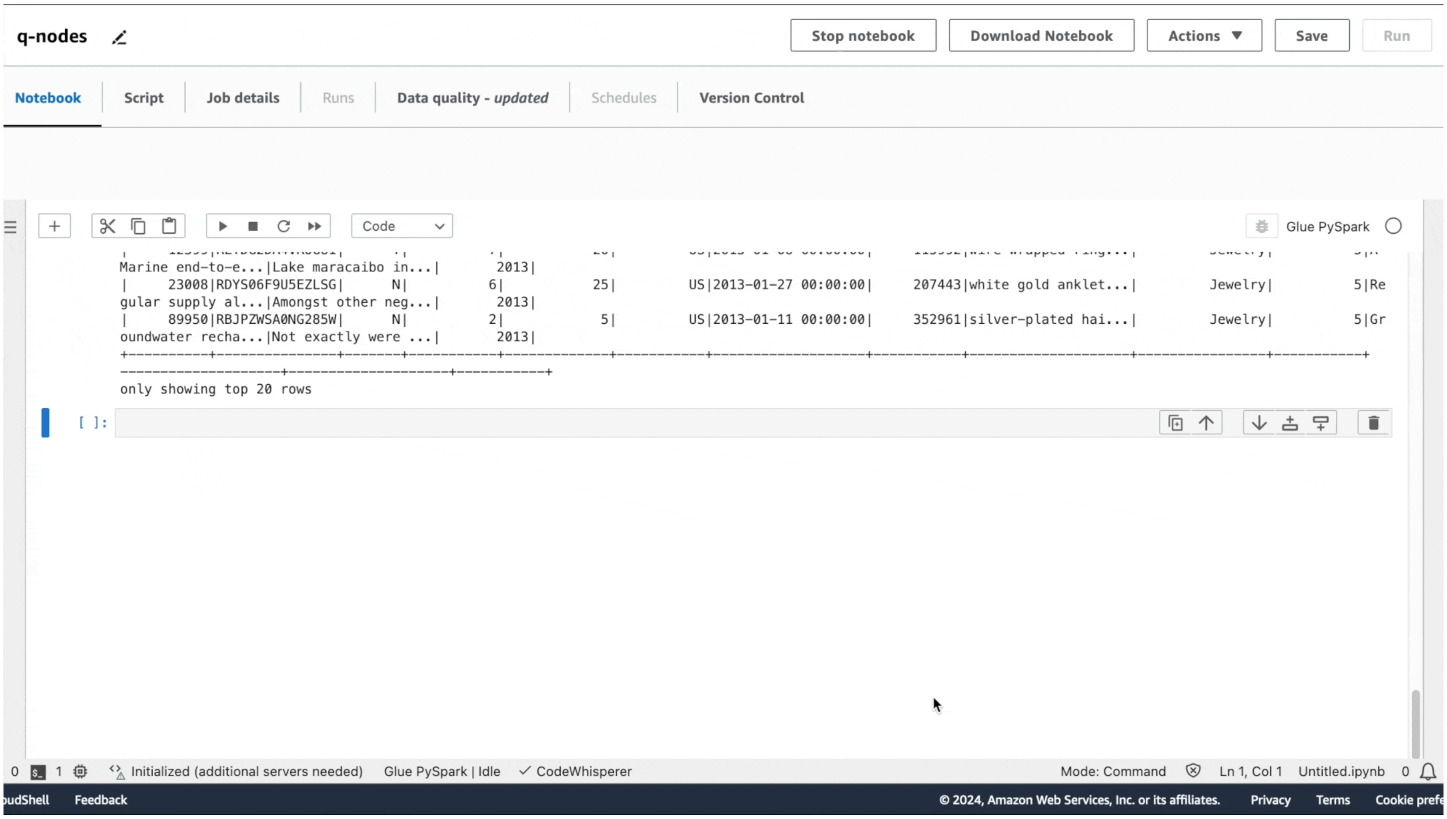

q-nodes
Stop notebook
Download Notebook
Actions ▼
Save
Run
Notebook
Script
Job details
Runs
Data quality - updated
Schedules
Version Control
Code
Glue PySpark ○

```
Worker Type: G.1X
Number of Workers: 5
Session ID: a6846a9a-6489-4599-bf8d-066b59d887da
Applying the following default arguments:
--glue_kernel_version 1.0.4
--enable-glue-datacatalog true
Waiting for session a6846a9a-6489-4599-bf8d-066b59d887da to get into ready status...
Session a6846a9a-6489-4599-bf8d-066b59d887da has been created.
```

[ ]:
0 1 Initialized (additional servers needed) Glue PySpark | Idle ✓ CodeWhisperer
Mode: Edit Ln 1, Col 1 Untitled.ipynb 0
loudShell Feedback
© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie pre


q-nodes
Stop notebook
Download Notebook
Actions ▼
Save
Run
Notebook
Script
Job details
Runs
Data quality - *updated*
Schedules
Version Control
Code
Glue PySpark
[ ]:
0 1 Initialized (additional servers needed) Glue PySpark | Idle ✓ CodeWhisperer
Mode: Edit Ln 1, Col 1 Untitled.ipynb 0
oudShell Feedback © 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie prefe

```
|       US|     171306|R00GN0TEQS4ISDM|     90211|white and yellow ...|          3|            5|          5|      N|PAP, and regular ...|Words themselves
...|2013-01-23 00:00:00|       2013|          Jewelry|
+-----------+-----------+---------------+----------+--------------------+-----------+-------------+-----------+-------+--------------------+
--------------------+--------------------+-----------+----------------+
only showing top 20 rows

/opt/amazon/spark/python/lib/pyspark.zip/pyspark/sql/dataframe.py:127: UserWarning: DataFrame constructor is internal. Do not directly use it.
```


q-nodes
Stop notebook
Download Notebook
Actions ▼
Save
Run
Notebook
Script
Job details
Runs
Data quality - *updated*
Schedules
Version Control
Code
Glue PySpark
[ ]:
0 1 Initialized (additional servers needed) Glue PySpark | Idle ✓ CodeWhisperer
Mode: Command Ln 1, Col 1 Untitled.ipynb 0
oudShell Feedback © 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie prefe

```
Marine end-to-e...|Lake maracaibo in...|       2013|
|     23008|RDYS06F9U5EZLSG|       N|          6|          25|          US|2013-01-27 00:00:00|     207443|white gold anklet...|          Jewelry|           5|Re
gular supply al...|Amongst other neg...|       2013|
|     89950|RBJPZWSA0NG285W|       N|          2|           5|          US|2013-01-11 00:00:00|     352961|silver-plated hai...|          Jewelry|           5|Gr
oundwater recha...|Not exactly were ...|       2013|
+----------+---------------+--------+-----------+------------+------------+-------------------+-----------+--------------------+-----------------+------------+
--------------------+--------------------+-----------+
only showing top 20 rows
```

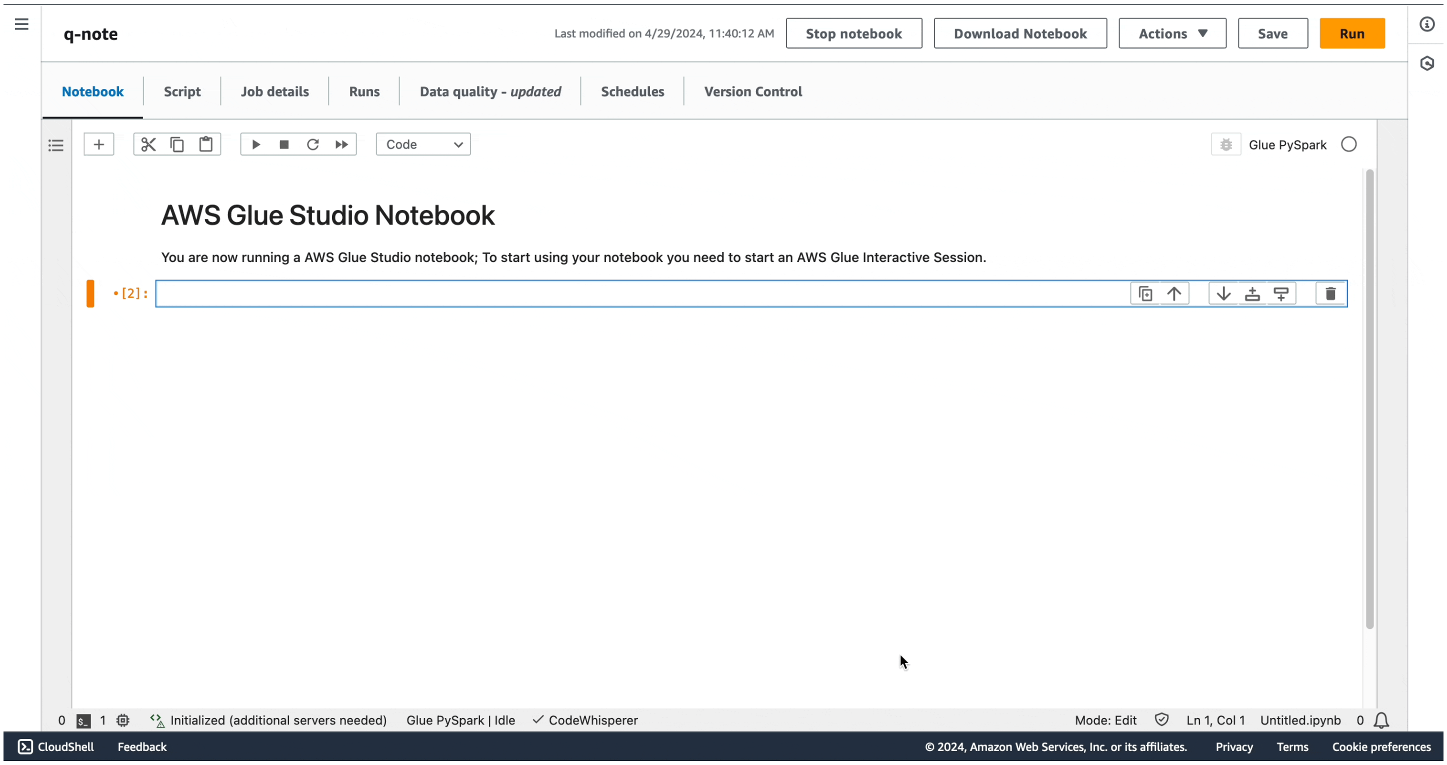
Similar to the Amazon Q chat experience, the code is recommended. If you press **Tab**, then the recommended code is chosen.

You can run each cell by filling in the appropriate options for your sources in the generated code. At any point in the runs, you can also preview a sample of your dataset by using the `show()` method.

You can run the notebook as a job, either programmatically or by choosing **Run**.

Complex prompts

You can generate a full script with a single complex prompt. *"I have JSON data in S3 and data in Oracle that needs combining. Please provide a Glue script that reads from both sources, does a join, and then writes results to Redshift."*



AWS Glue Studio Notebook

You are now running a AWS Glue Studio notebook; To start using your notebook you need to start an AWS Glue Interactive Session.

q-note
Last modified on 4/29/2024, 11:40:12 AM
Stop notebook
Download Notebook
Actions ▼
Save
Run
Notebook
Script
Job details
Runs
Data quality - updated
Schedules
Version Control
Code
Glue PySpark
[2]:
0 1 Initialized (additional servers needed)
Glue PySpark | Idle
✓ CodeWhisperer
Mode: Edit
Ln 1, Col 1
Untitled.ipynb
0
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

You may notice that, on the notebook, Amazon Q data integration in AWS Glue generated the same code snippet that was generated in the Amazon Q chat.

You can run the notebook as a job, either by choosing **Run** or programmatically.

Using context awareness with Amazon Q Data Integration in AWS Glue

You can now create data processing jobs more efficiently with query based context-aware and PySpark DataFrame code generation in Amazon Q Data Integration. For example, you can use this prompt to generate PySpark code: “create a job to load sales data from Redshift table ‘analytics.salesorder’ using connection ‘erp_conn’, filter order_amount below 50 dollars, and save to Amazon S3 in parquet format.”

Amazon Q will generate the script based on your prompt and setup data integration workflow setup with the details provided from your question, such as connection configurations, schema details, database/table names, and column specifications for transformations. Sensitive information, such as connection option passwords, continue to be redacted.

If required information is not provided from your prompt, Amazon Q will put placeholders, which you have to update the generated code with the appropriate values before running the code.

The following are examples on how to use context awareness.

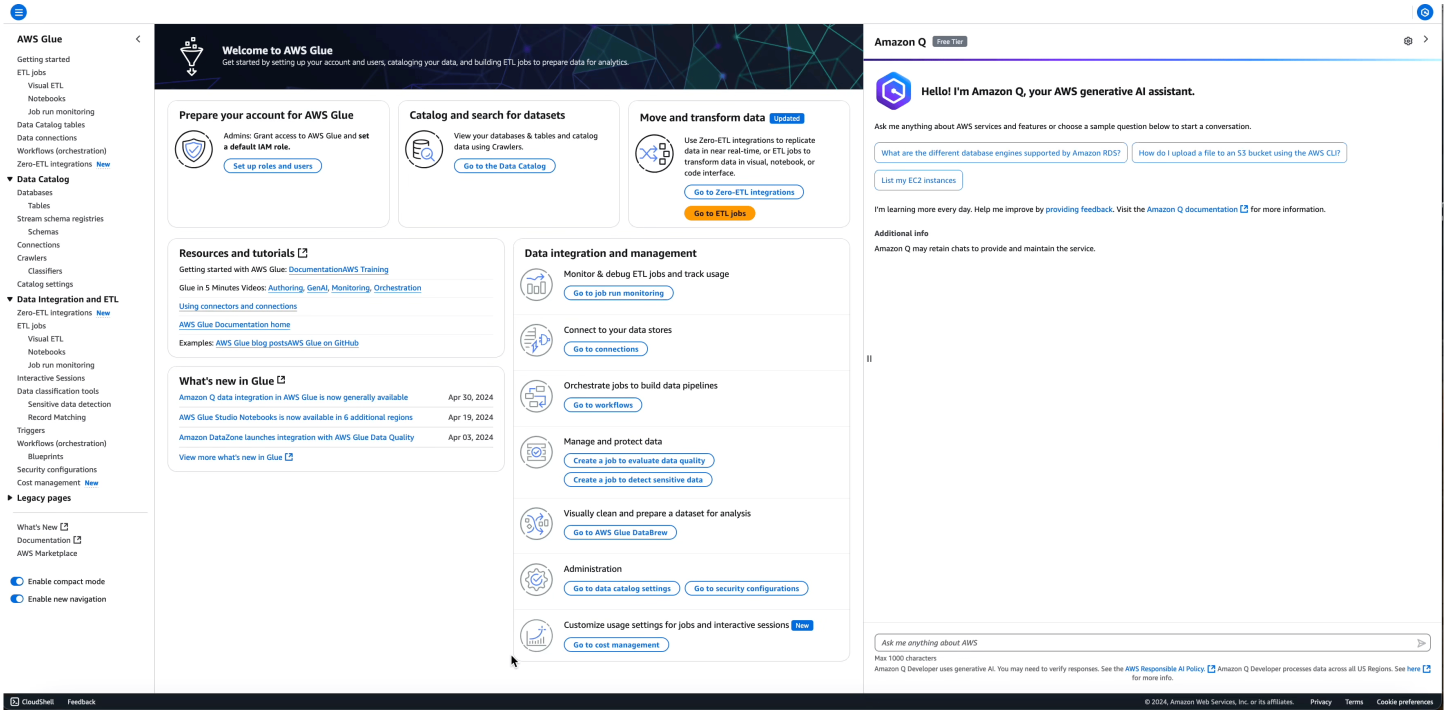
Example: interactions

Reading an Amazon S3 file and writing to an Amazon S3 bucket

Prompt: `Create an AWS Glue spark job that reads a file s3://amzn-s3-demo-bucket-input-bucket-name/input-filename.csv and saves it into the parquet file under directory s3://amzn-s3-demo-bucket-output-bucket-name/output-s3-prefix/`

```
import sys
from awsglue.transforms import *
from pyspark.context import SparkContext
from pyspark.sql import SparkSession

sc = SparkContext.getOrCreate()
spark = SparkSession.builder.getOrCreate()

# Script generated for node S3DataSource
S3DataSource_dsource1 = spark.read.format("csv").load(
    "s3://amzn-s3-demo-bucket-input-bucket-name/input-filename.csv"
)
```

```
# Script generated for node S3DataSink
S3DataSource_dsource1.write.format("parquet").mode("append").save(
    "s3://amzn-s3-demo-bucket-output-bucket-name/output-s3-prefix/"
)
```



AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Zero-ETL integrations New
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Zero-ETL integrations New
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Interactive Sessions
Data classification tools
Sensitive data detection
Record Matching
Triggers
Workflows (orchestration)
Blueprints
Security configurations
Cost management New
Legacy pages
What's New
Documentation
AWS Marketplace
Enable compact mode
Enable new navigation

Welcome to AWS Glue

Get started by setting up your account and users, cataloging your data, and building ETL jobs to prepare data for analytics.

Prepare your account for AWS Glue

Admins: Grant access to AWS Glue and **set a default IAM role.**

Set up roles and users

Catalog and search for datasets

View your databases & tables and catalog data using Crawlers.

Go to the Data Catalog

Move and transform data Updated

Use Zero-ETL integrations to replicate data in near real-time, or ETL jobs to transform data in visual, notebook, or code interface.

Go to Zero-ETL integrations
Go to ETL jobs

Resources and tutorials

Getting started with AWS Glue: DocumentationAWS Training

Glue in 5 Minutes Videos: Authoring, GenAI, Monitoring, Orchestration

Using connectors and connections

AWS Glue Documentation home

Examples: AWS Glue blog postsAWS Glue on GitHub

What's new in Glue

Amazon Q data integration in AWS Glue is now generally available	Apr 30, 2024
AWS Glue Studio Notebooks is now available in 6 additional regions	Apr 19, 2024
Amazon DataZone launches integration with AWS Glue Data Quality	Apr 03, 2024

View more what's new in Glue

Data integration and management

Monitor & debug ETL jobs and track usage

Go to job run monitoring

Connect to your data stores

Go to connections

Orchestrate jobs to build data pipelines

Go to workflows

Manage and protect data

Create a job to evaluate data quality
Create a job to detect sensitive data

Visually clean and prepare a dataset for analysis

Go to AWS Glue DataBrew

Administration

Go to data catalog settings
Go to security configurations

Customize usage settings for jobs and interactive sessions New

Go to cost management

Amazon Q Free Tier

Hello! I'm Amazon Q, your AWS generative AI assistant.

Ask me anything about AWS services and features or choose a sample question below to start a conversation.

What are the different database engines supported by Amazon RDS?
How do I upload a file to an S3 bucket using the AWS CLI?
List my EC2 instances

I'm learning more every day. Help me improve by providing feedback. Visit the Amazon Q documentation for more information.

Additional info

Amazon Q may retain chats to provide and maintain the service.

Ask me anything about AWS

Max 1000 characters

Amazon Q Developer uses generative AI. You may need to verify responses. See the AWS Responsible AI Policy. Amazon Q Developer processes data across all US Regions. See here for more info.

CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

Get data from Lakehouse and write to database

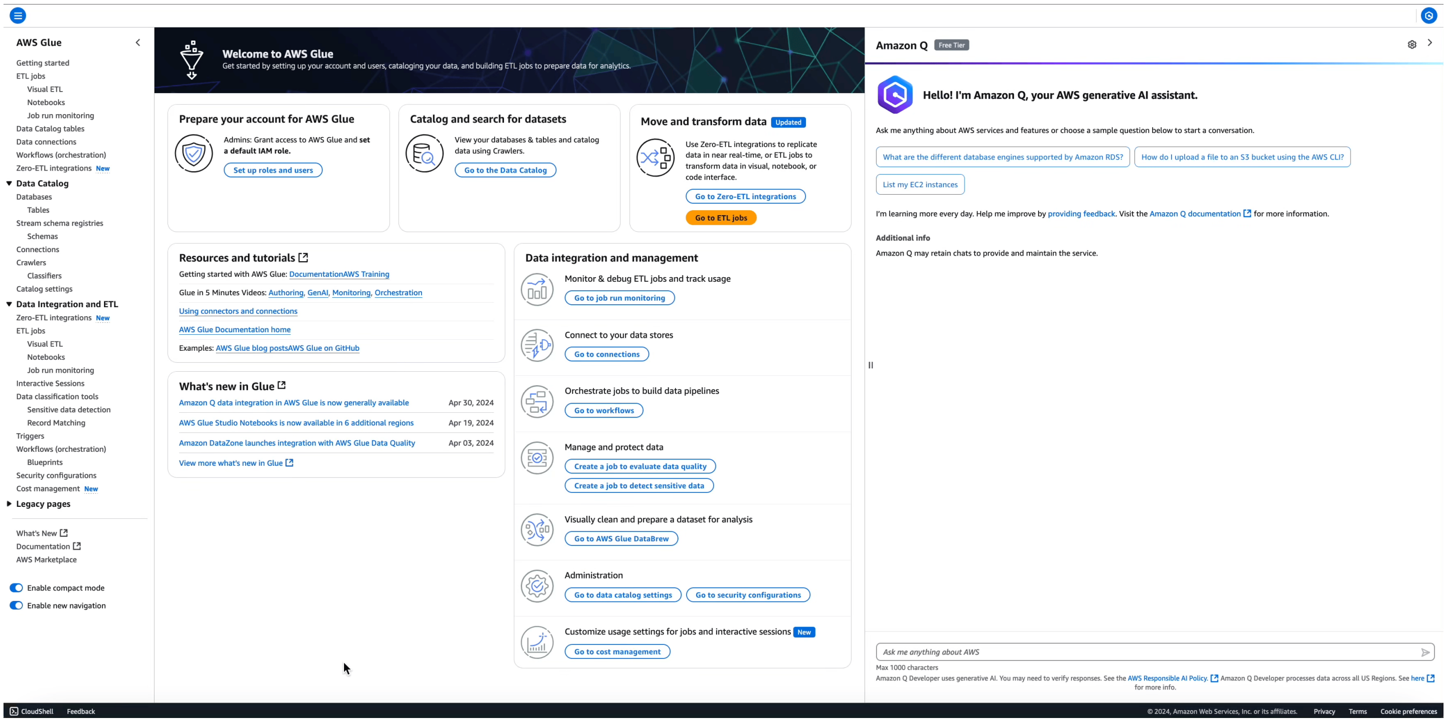
Prompt: `write an ETL script to read from a Lakehouse table my-table in database my-database and write it to a RDS MySQL table my-target-table`

For fields where you did not provide information (for example, the connectionName required is for MySQL data sink and default with a placehoder <connection-name> in the generated code), a placeholder is kept for you to fill in the required information before run the script.

Generated script:

```
import sys
from awsglue.transforms import *
from pyspark.context import SparkContext
from pyspark.sql import SparkSession
from connectivity.adapter import CatalogConnectionHelper

sc = SparkContext.getOrCreate()
```

```
spark = SparkSession.builder.getOrCreate()

# Script generated for node S3DataSource
S3DataSource_dsource1 = spark.read.format("parquet").load(
    "s3://amzn-lakehouse-demo-bucket/my-database/my-table"
)

# Script generated for node ConnectionV2DataSink
ConnectionV2DataSink_dsink1_additional_options = {"dbtable": "my-target-table"}
CatalogConnectionHelper(spark).write(
    S3DataSource_dsource1,
    "mysql",
    "<connection-name>",
    ConnectionV2DataSink_dsink1_additional_options,
)
```


AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Zero-ETL integrations New
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Zero-ETL integrations New
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Interactive Sessions
Data classification tools
Sensitive data detection
Record Matching
Triggers
Workflows (orchestration)
Blueprints
Security configurations
Cost management New
Legacy pages
What's New
Documentation
AWS Marketplace
Enable compact mode
Enable new navigation
Set up roles and users
Go to the Data Catalog
Updated
Go to Zero-ETL integrations
Go to ETL jobs
Go to job run monitoring
Go to connections
Go to workflows
Create a job to evaluate data quality
Create a job to detect sensitive data
Go to AWS Glue DataBrew
Go to data catalog settings
Go to security configurations
New
Go to cost management
Free Tier
What are the different database engines supported by Amazon RDS?
How do I upload a file to an S3 bucket using the AWS CLI?
List my EC2 instances
Ask me anything about AWS
CloudShell
Feedback
Privacy
Terms
Cookie preferences

Welcome to AWS Glue

Get started by setting up your account and users, cataloging your data, and building ETL jobs to prepare data for analytics.

Prepare your account for AWS Glue

Admins: Grant access to AWS Glue and **set a default IAM role.**

Catalog and search for datasets

View your databases & tables and catalog data using Crawlers.

Move and transform data

Use Zero-ETL integrations to replicate data in near real-time, or ETL jobs to transform data in visual, notebook, or code interface.

Resources and tutorials

Getting started with AWS Glue: DocumentationAWS Training

Glue in 5 Minutes Videos: Authoring, GenAI, Monitoring, Orchestration

Using connectors and connections

AWS Glue Documentation home

Examples: AWS Glue blog postsAWS Glue on GitHub

What's new in Glue

Amazon Q data integration in AWS Glue is now generally available	Apr 30, 2024
AWS Glue Studio Notebooks is now available in 6 additional regions	Apr 19, 2024
Amazon DataZone launches integration with AWS Glue Data Quality	Apr 03, 2024

View more what's new in Glue

Data integration and management

Monitor & debug ETL jobs and track usage

Connect to your data stores

Orchestrate jobs to build data pipelines

Manage and protect data

Visually clean and prepare a dataset for analysis

Administration

Customize usage settings for jobs and interactive sessions

Amazon Q

Hello! I'm Amazon Q, your AWS generative AI assistant.

Ask me anything about AWS services and features or choose a sample question below to start a conversation.

I'm learning more every day. Help me improve by providing feedback. Visit the Amazon Q documentation for more information.

Additional info

Amazon Q may retain chats to provide and maintain the service.

Max 1000 characters

Amazon Q Developer uses generative AI. You may need to verify responses. See the AWS Responsible AI Policy. Amazon Q Developer processes data across all US Regions. See here for more info.

© 2024, Amazon Web Services, Inc. or its affiliates.

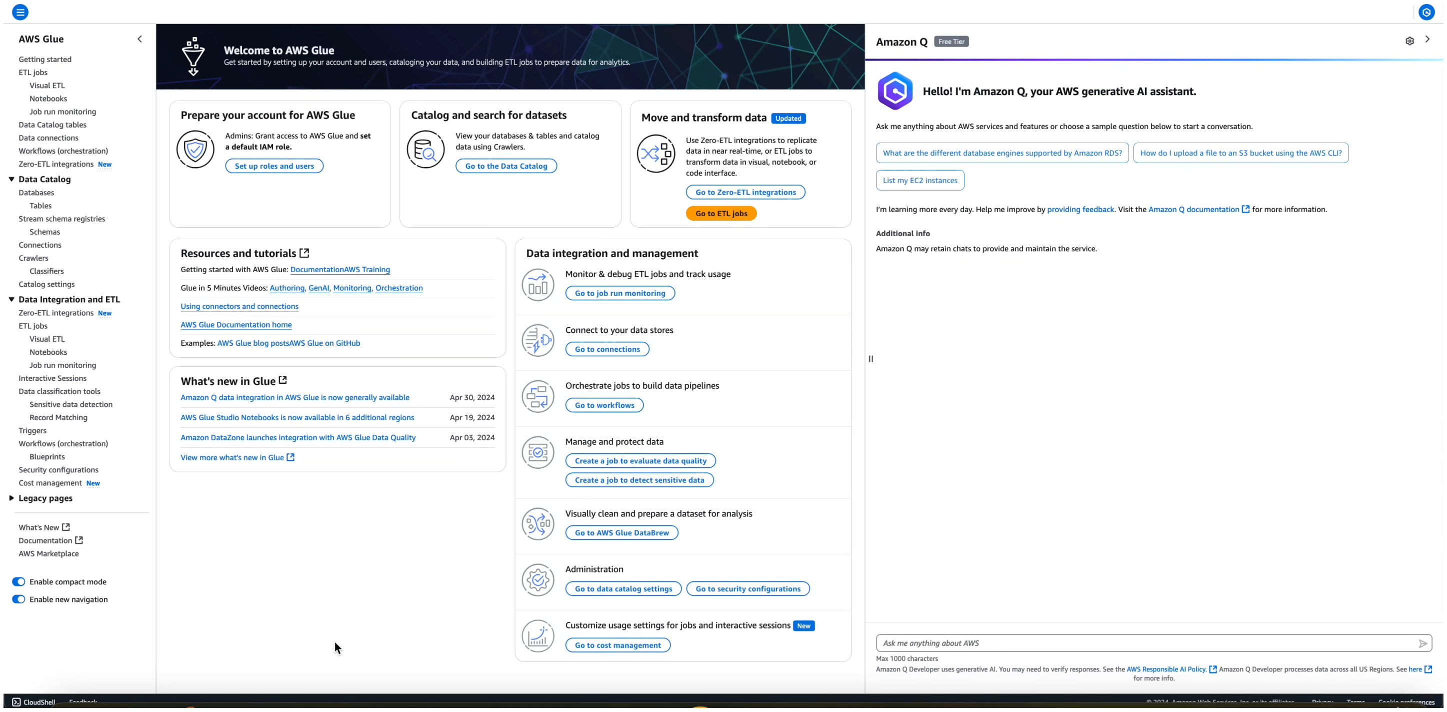
Example: Full ETL workflow

The following example demonstrates how you can ask AWS Glue to create a AWS Glue script to complete a full ETL workflow with the following prompt: `Create a AWS Glue ETL Script read from two AWS Glue Data Catalog tables venue and event in my database glue_db_4fthqih3vvk1if, join the results on the field venueid, filter on venue state with condition as venuestate=='DC' after joining the results` `and write output to an Amazon S3 S3 location s3://amz-s3-demo-bucket/ output/ in CSV format`.

The workflow contains reading from different data sources (two AWS Glue Data Catalog tables), and a couple of transforms after the reading by join the result from two readings, filter based on some condition and write the transformed output to an Amazon S3 destination in CSV format.

The generated job will fill in the detailed information to the data source, transform and sink operation with corresponding information extracted from user question as below.

```
import sys
from awsglue.transforms import *
from pyspark.context import SparkContext
from pyspark.sql import SparkSession

sc = SparkContext.getOrCreate()
spark = SparkSession.builder.getOrCreate()

# Script generated for node CatalogDataSource
CatalogDataSource_dsource1 = spark.sql("select * from
 `glue_db_4fthqih3vvk1if`.`venue`")

# Script generated for node CatalogDataSource
CatalogDataSource_dsource2 = spark.sql("select * from
 `glue_db_4fthqih3vvk1if`.`event`")

# Script generated for node JoinTransform
JoinTransform_transform1 = CatalogDataSource_dsource1.join(
    CatalogDataSource_dsource2,
    (CatalogDataSource_dsource1["venueid"] == CatalogDataSource_dsource2["venueid"]),
    "inner",
)

# Script generated for node FilterTransform
FilterTransform_transform2 = JoinTransform_transform1.filter("venuestate=='DC'")

# Script generated for node S3DataSink
FilterTransform_transform2.write.format("csv").mode("append").save(
    "s3://amz-s3-demo-bucket/output//output/"
)
```


AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Zero-ETL integrations New
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Zero-ETL integrations New
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Interactive Sessions
Data classification tools
Sensitive data detection
Record Matching
Triggers
Workflows (orchestration)
Blueprints
Security configurations
Cost management New
Legacy pages
What's New
Documentation
AWS Marketplace
Enable compact mode
Enable new navigation
Welcome to AWS Glue
Prepare your account for AWS Glue
Set up roles and users
Catalog and search for datasets
Go to the Data Catalog
Move and transform data Updated
Go to Zero-ETL integrations
Go to ETL jobs
Resources and tutorials
Getting started with AWS Glue: DocumentationAWS Training
Glue in 5 Minutes Videos: Authoring, GenAI, Monitoring, Orchestration
Using connectors and connections
AWS Glue Documentation home
Examples: AWS Glue blog postsAWS Glue on GitHub
What's new in Glue
Amazon Q data integration in AWS Glue is now generally available Apr 30, 2024
AWS Glue Studio Notebooks is now available in 6 additional regions Apr 19, 2024
Amazon DataZone launches integration with AWS Glue Data Quality Apr 03, 2024
View more what's new in Glue
Data integration and management
Monitor & debug ETL jobs and track usage
Go to job run monitoring
Connect to your data stores
Go to connections
Orchestrate jobs to build data pipelines
Go to workflows
Manage and protect data
Create a job to evaluate data quality
Create a job to detect sensitive data
Visually clean and prepare a dataset for analysis
Go to AWS Glue DataBrew
Administration
Go to data catalog settings
Go to security configurations
Customize usage settings for jobs and interactive sessions New
Go to cost management
Amazon Q Free Tier
What are the different database engines supported by Amazon RDS?
How do I upload a file to an S3 bucket using the AWS CLI?
List my EC2 instances
Ask me anything about AWS
CloudShell
Feedback

Get started by setting up your account and users, cataloging your data, and building ETL jobs to prepare data for analytics.

Admins: Grant access to AWS Glue and **set a default IAM role.**

View your databases & tables and catalog data using Crawlers.

Use Zero-ETL integrations to replicate data in near real-time, or ETL jobs to transform data in visual, notebook, or code interface.

Hello! I'm Amazon Q, your AWS generative AI assistant.

Ask me anything about AWS services and features or choose a sample question below to start a conversation.

I'm learning more every day. Help me improve by providing feedback. Visit the Amazon Q documentation for more information.

Additional info

Amazon Q may retain chats to provide and maintain the service.

Max 1000 characters

Amazon Q Developer uses generative AI. You may need to verify responses. See the AWS Responsible AI Policy. Amazon Q Developer processes data across all US Regions. See here for more info.

Limitations

- Context carryover:
  - The context-awareness feature only carries over context from the previous user query within the same conversation. It does not retain context beyond the immediate preceding query.
- Support for node configurations:
  - Currently, context-awareness supports only a subset of required configurations for various nodes.
  - Support for optional fields is planned in upcoming releases.
- Availability:
  - Context-awareness and DataFrame support are available in Q Chat and SageMaker Unified Studio notebooks. However, these features are not yet available in AWS Glue Studio notebooks.

Orchestration in AWS Glue

The following sections provide information on orchestration of jobs in AWS Glue.

Topics

- Starting jobs and crawlers using triggers
- Performing complex ETL activities using blueprints and workflows in AWS Glue
- Developing blueprints in AWS Glue

Starting jobs and crawlers using triggers

In AWS Glue, you can create Data Catalog objects called triggers, which you can use to either manually or automatically start one or more crawlers or extract, transform, and load (ETL) jobs. Using triggers, you can design a chain of dependent jobs and crawlers.

> **Note**
>
> You can accomplish the same thing by defining *workflows*. Workflows are preferred for creating complex multi-job ETL operations. For more information, see the section called "Performing complex ETL activities using blueprints and workflows".

Topics

- AWS Glue triggers
- Adding triggers
- Activating and deactivating triggers

AWS Glue triggers

When *fired*, a trigger can start specified jobs and crawlers. A trigger fires on demand, based on a schedule, or based on a combination of events.

Note

Only two crawlers can be activated by a single trigger. If you want to crawl multiple data stores, use multiple sources for each crawler instead of running multiple crawlers simultaneously.

A trigger can exist in one of several states. A trigger is either `CREATED`, `ACTIVATED`, or `DEACTIVATED`. There are also transitional states, such as `ACTIVATING`. To temporarily stop a trigger from firing, you can deactivate it. You can then reactivate it later.

There are three types of triggers:

Scheduled

A time-based trigger based on `cron`.

You can create a trigger for a set of jobs or crawlers based on a schedule. You can specify constraints, such as the frequency that the jobs or crawlers run, which days of the week they run, and at what time. These constraints are based on `cron`. When you're setting up a schedule for a trigger, consider the features and limitations of cron. For example, if you choose to run your crawler on day 31 each month, keep in mind that some months don't have 31 days. For more information about cron, see Time-based schedules for jobs and crawlers.

Conditional

A trigger that fires when a previous job or crawler or multiple jobs or crawlers satisfy a list of conditions.

When you create a conditional trigger, you specify a list of jobs and a list of crawlers to watch. For each watched job or crawler, you specify a status to watch for, such as succeeded, failed, timed out, and so on. The trigger fires if the watched jobs or crawlers end with the specified statuses. You can configure the trigger to fire when any or all of the watched events occur.

For example, you could configure a trigger T1 to start job J3 when both job J1 and job J2 successfully complete, and another trigger T2 to start job J4 if either job J1 or job J2 fails.

The following table lists the job and crawler completion states (events) that triggers watch for.

Job completion states	Crawler completion states

`SUCCEEDED`
`STOPPED`
`FAILED`
`TIMEOUT`
	`SUCCEEDED`
	`FAILED`
	`CANCELLED`

On-demand

A trigger that fires when you activate it. On-demand triggers never enter the `ACTIVATED` or `DEACTIVATED` state. They always remain in the `CREATED` state.

So that they are ready to fire as soon as they exist, you can set a flag to activate scheduled and conditional triggers when you create them.

> ⚠ **Important**
>
> Jobs or crawlers that run as a result of other jobs or crawlers completing are referred to as *dependent*. Dependent jobs or crawlers are only started if the job or crawler that completes was started by a trigger. All jobs or crawlers in a dependency chain must be descendants of a single **scheduled** or **on-demand** trigger.

Passing job parameters with triggers

A trigger can pass parameters to the jobs that it starts. Parameters include job arguments, timeout value, security configuration, and more. If the trigger starts multiple jobs, the parameters are passed to each job.

The following are the rules for job arguments passed by a trigger:

- If the key in the key-value pair matches a default job argument, the passed argument overrides the default argument. If the key doesn’t match a default argument, then the argument is passed as an additional argument to the job.
- If the key in the key-value pair matches a non-overridable argument, the passed argument is ignored.

For more information, see the section called "Triggers" in the AWS Glue API.

Adding triggers

You can add a trigger using the AWS Glue console, the AWS Command Line Interface (AWS CLI), or the AWS Glue API.

> **Note**
>
> Currently, the AWS Glue console supports only jobs, not crawlers, when working with triggers. You can use the AWS CLI or AWS Glue API to configure triggers with both jobs and crawlers.

To add a trigger (console)

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, under **ETL**, choose **Triggers**. Then choose **Add trigger**.
3. Provide the following properties:

   Name

   Give your trigger a unique name.

   Trigger type

   Specify one of the following:

   - **Schedule:** The trigger fires at a specific frequency and time.
   - **Job events:** A conditional trigger. The trigger fires when any or all jobs in the list match their designated statuses. For the trigger to fire, the watched jobs must have been started by triggers. For any job you choose, you can only watch one job event (completion status).
   - **On-demand:** The trigger fires when it is activated.
4. Complete the trigger wizard. On the **Review** page, you can activate **Schedule** and **Job events** (conditional) triggers immediately by selecting **Enable trigger on creation**.

To add a trigger (AWS CLI)

- Enter a command similar to the following.

  ```
  aws glue create-trigger --name MyTrigger --type SCHEDULED --schedule  "cron(0 12 *
   * ? *)" --actions CrawlerName=MyCrawler --start-on-creation
  ```

  This command creates a schedule trigger named `MyTrigger`, which runs every day at 12:00pm UTC and starts a crawler named `MyCrawler`. The trigger is created in the activated state.

For more information, see the section called “AWS Glue triggers”.

Time-based schedules for jobs and crawlers

You can define a time-based schedule for your crawlers and jobs in AWS Glue. The definition of these schedules uses the Unix-like cron syntax. You specify time in Coordinated Universal Time (UTC), and the minimum precision for a schedule is 5 minutes.

To learn more about configuring jobs and crawlers to run using a schedule, see Starting jobs and crawlers using triggers.

Cron expressions

Cron expressions have six required fields, which are separated by white space.

Syntax

```
cron(Minutes Hours Day-of-month Month Day-of-week Year)
```

Fields	Values	Wildcards
Minutes	0–59	, - * /
Hours	0–23	, - * /
Day-of-month	1–31	, - * ? / L W
Month	1–12 or JAN-DEC	, - * /

Fields	Values	Wildcards
Day-of-week	1–7 or SUN-SAT	, - * ? / L
Year	1970–2199	, - * /

Wildcards

- The **,** (comma) wildcard includes additional values. In the `Month` field, `JAN,FEB,MAR` would include January, February, and March.
- The **-** (dash) wildcard specifies ranges. In the `Day` field, 1–15 would include days 1 through 15 of the specified month.
- The ***** (asterisk) wildcard includes all values in the field. In the `Hours` field, ***** would include every hour.
- The **/** (forward slash) wildcard specifies increments. In the `Minutes` field, you could enter **`1/10`** to specify every 10th minute, starting from the first minute of the hour (for example, the 11th, 21st, and 31st minute).
- The **?** (question mark) wildcard specifies one or another. In the `Day-of-month` field you could enter **7**, and if you didn't care what day of the week the seventh was, you could enter **?** in the Day-of-week field.
- The **L** wildcard in the `Day-of-month` or `Day-of-week` fields specifies the last day of the month or week.
- The **W** wildcard in the `Day-of-month` field specifies a weekday. In the `Day-of-month` field, `3W` specifies the day closest to the third weekday of the month.

Limits

- You can't specify the `Day-of-month` and `Day-of-week` fields in the same cron expression. If you specify a value in one of the fields, you must use a **?** (question mark) in the other.
- Cron expressions that lead to rates faster than 5 minutes are not supported.

Examples

When creating a schedule, you can use the following sample cron strings.

Minutes	Hours	Day of month	Month	Day of week	Year	Meaning
0	10	*	*	?	*	Run at 10:00 am (UTC) every day
15	12	*	*	?	*	Run at 12:15 pm (UTC) every day
0	18	?	*	MON-FRI	*	Run at 6:00 pm (UTC) every Monday through Friday
0	8	1	*	?	*	Run at 8:00 am (UTC) every first day of the month
0/15	*	*	*	?	*	Run every 15 minutes
0/10	*	?	*	MON-FRI	*	Run every 10 minutes Monday through Friday

Minutes	Hours	Day of month	Month	Day of week	Year	Meaning
0/5	8–17	?	*	MON-FRI	*	Run every 5 minutes Monday through Friday between 8:00 am and 5:55 pm (UTC)

For example to run on a schedule of every day at 12:15 UTC, specify:

```
cron(15 12 * * ? *)
```

Activating and deactivating triggers

You can activate or deactivate a trigger using the AWS Glue console, the AWS Command Line Interface (AWS CLI), or the AWS Glue API.

To activate or deactivate a trigger (console)

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, under **ETL**, choose **Triggers**.
3. Select the check box next to the desired trigger, and on the **Action** menu choose **Enable trigger** to activate the trigger or **Disable trigger** to deactivate the trigger.

To activate or deactivate a trigger (AWS CLI)

- Enter one of the following commands.

  ```
  aws glue start-trigger --name MyTrigger
  ```

```
aws glue stop-trigger --name MyTrigger
```

Starting a trigger activates it, and stopping a trigger deactivates it. When you activate an on-demand trigger, it fires immediately.

For more information, see the section called “AWS Glue triggers”.

Performing complex ETL activities using blueprints and workflows in AWS Glue

Some of your organization's complex extract, transform, and load (ETL) processes might best be implemented by using multiple, dependent AWS Glue jobs and crawlers. Using AWS Glue *workflows*, you can design a complex multi-job, multi-crawler ETL process that AWS Glue can run and track as single entity. After you create a workflow and specify the jobs, crawlers, and triggers in the workflow, you can run the workflow on demand or on a schedule.

Topics

- Overview of workflows in AWS Glue
- Creating and building out a workflow manually in AWS Glue
- Starting an AWS Glue workflow with an Amazon EventBridge event
- Viewing the EventBridge events that started a workflow
- Running and monitoring a workflow in AWS Glue
- Stopping a workflow run
- Repairing and resuming a workflow run
- Getting and setting workflow run properties in AWS Glue
- Querying workflows using the AWS Glue API
- Blueprint and workflow restrictions in AWS Glue
- Troubleshooting blueprint errors in AWS Glue
- Permissions for personas and roles for AWS Glue blueprints

Overview of workflows in AWS Glue

In AWS Glue, you can use workflows to create and visualize complex extract, transform, and load (ETL) activities involving multiple crawlers, jobs, and triggers. Each workflow manages the execution and monitoring of all its jobs and crawlers. As a workflow runs each component, it records execution progress and status. This provides you with an overview of the larger task and the details of each step. The AWS Glue console provides a visual representation of a workflow as a graph.

You can create a workflow from an AWS Glue blueprint, or you can manually build a workflow a component at a time using the AWS Management Console or the AWS Glue API. For more information about blueprints, see the section called "Overview of blueprints".

Triggers within workflows can start both jobs and crawlers and can be fired when jobs or crawlers complete. By using triggers, you can create large chains of interdependent jobs and crawlers. In addition to triggers within a workflow that define job and crawler dependencies, each workflow has a *start trigger*. There are three types of start triggers:

- **Schedule –** The workflow is started according to a schedule that you define. The schedule can be daily, weekly, monthly, and so on, or can be a custom schedule based on a `cron` expression.
- **On demand –** The workflow is started manually from the AWS Glue console, API, or AWS CLI.
- **EventBridge event –** The workflow is started upon the occurrence of a single Amazon EventBridge event or a batch of Amazon EventBridge events. With this trigger type, AWS Glue can be an event consumer in an event-driven architecture. Any EventBridge event type can start a workflow. A common use case is the arrival of a new object in an Amazon S3 bucket (the S3 `PutObject` operation).

  Starting a workflow with a batch of events means waiting until a specified number of events have been received or until a specified amount of time has passed. When you create the EventBridge event trigger, you can optionally specify batch conditions. If you specify batch conditions, you must specify the batch size (number of events), and can optionally specify a batch window (number of seconds). The default and maximum batch window is 900 seconds (15 minutes). The batch condition that is met first starts the workflow. The batch window starts when the first event arrives. If you don't specify batch conditions when creating a trigger, the batch size defaults to 1.

  When the workflow starts, the batch conditions are reset and the event trigger begins watching for the next batch condition to be met to start the workflow again.

  The following table shows how batch size and batch window operate together to trigger a workflow.

Batch size	Batch window	Resulting triggering condition
10		The workflow is triggered upon the arrival of 10 EventBridge events, or 15 minutes after the arrival of the first event, whichever occurs first. (If windows size isn't specified, it defaults to 15 minutes.)
10	2 mins	The workflow is triggered upon the arrival of 10 EventBridge events, or 2 minutes after the arrival of the first event, whichever occurs first.
1		The workflow is triggered upon the arrival of the first event. Window size is irrelevant. The batch size defaults to 1 if you don't specify batch conditions when you create the EventBridge event trigger.

The `GetWorkflowRun` API operation returns the batch condition that triggered the workflow.

Regardless of how a workflow is started, you can specify the maximum number of concurrent workflow runs when you create the workflow.

If an event or batch of events starts a workflow run that eventually fails, that event or batch of events is no longer considered for starting a workflow run. A new workflow run is started only when the next event or batch of events arrives.

> ⚠ **Important**
>
> Limit the total number of jobs, crawlers, and triggers within a workflow to 100 or less. If you include more than 100, you might get errors when trying to resume or stop workflow runs.

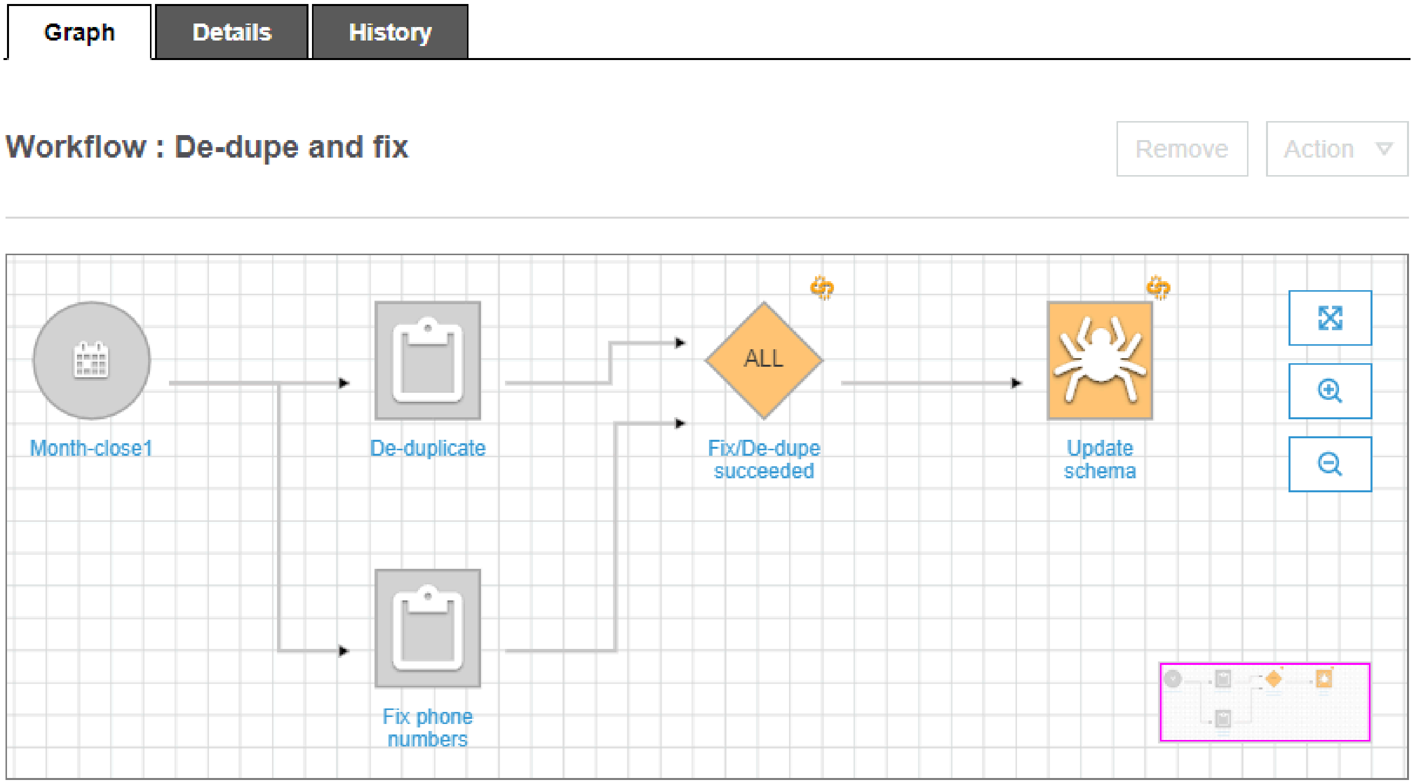
A workflow run will not be started if it would exceed the concurrency limit set for the workflow, even though the event condition is met. It's advisable to adjust workflow concurrency limits based on the expected event volume. AWS Glue does not retry workflow runs that fail due to exceeded concurrency limits. Likewise, it's advisable to adjust concurrency limits for jobs and crawlers within workflows based on expected event volume.

Workflow run properties

To share and manage state throughout a workflow run, you can define default workflow run properties. These properties, which are name/value pairs, are available to all the jobs in the workflow. Using the AWS Glue API, jobs can retrieve the workflow run properties and modify them for jobs that come later in the workflow.

Workflow graph

The following image shows the graph of a very basic workflow on the AWS Glue console. Your workflow could have dozens of components.


Graph
Details
History
Workflow : De-dupe and fix
Remove
Action
Month-close1
De-duplicate
ALL
Fix/De-dupe succeeded
Update schema
Fix phone numbers

This workflow is started by a schedule trigger, `Month-close1`, which starts two jobs, `De-duplicate` and `Fix phone numbers`. Upon successful completion of both jobs, an event trigger, `Fix/De-dupe succeeded`, starts a crawler, `Update schema`.

Static and dynamic workflow views

For each workflow, there is the notion of *static view* and *dynamic view*. The static view indicates the design of the workflow. The dynamic view is a runtime view that includes the latest run information for each of the jobs and crawlers. Run information includes success status and error details.

When a workflow is running, the console displays the dynamic view, graphically indicating the jobs that have completed and that are yet to be run. You can also retrieve a dynamic view of a running workflow using the AWS Glue API. For more information, see Querying workflows using the AWS Glue API.

See also

- the section called "Creating a workflow from a blueprint"
- the section called "Creating and building out a workflow manually"
- Workflows (for the workflows API)

Creating and building out a workflow manually in AWS Glue

You can use the AWS Glue console to manually create and build out a workflow one node at a time.

A workflow contains jobs, crawlers, and triggers. Before manually creating a workflow, create the jobs and crawlers that the workflow is to include. It's best to specify run-on-demand crawlers for workflows. You can create new triggers while you are building out your workflow, or you can *clone* existing triggers into the workflow. When you clone a trigger, all the catalog objects associated with the trigger—the jobs or crawlers that fire it and the jobs or crawlers that it starts—are added to the workflow.

Important

Limit the total number of jobs, crawlers, and triggers within a workflow to 100 or less. If you include more than 100, you might get errors when trying to resume or stop workflow runs.

You build out your workflow by adding triggers to the workflow graph, and defining the watched events and actions for each trigger. You begin with a *start trigger*, which can be either an on-demand or schedule trigger, and complete the graph by adding event (conditional) triggers.

Step 1: Create the workflow

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, under **ETL**, choose **Workflows**.
3. Choose **Add workflow** and complete the **Add a new ETL workflow** form.

   Any optional default run properties that you add are made available as arguments to all jobs in the workflow. For more information, see Getting and setting workflow run properties in AWS Glue.
4. Choose **Add workflow**.

   The new workflow appears in the list on the **Workflows** page.

Step 2: Add a start trigger

1. On the **Workflows** page, select your new workflow. Then, at the bottom of the page, ensure that the **Graph** tab is selected.
2. Choose **Add trigger**, and in the **Add trigger** dialog box, do one of the following:
   - Choose **Clone existing**, and choose a trigger to clone. Then choose **Add**.

     The trigger appears on the graph, along with the jobs and crawlers that it watches and the jobs and crawlers that it starts.

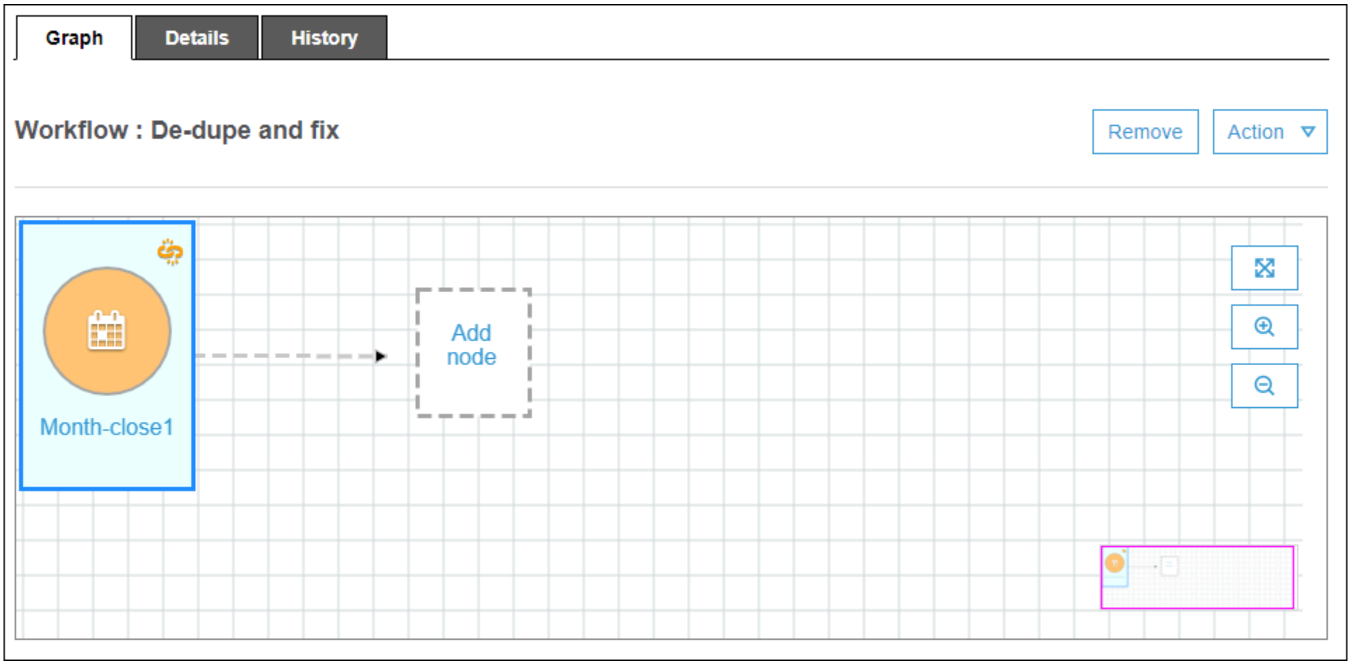
     If you mistakenly selected the wrong trigger, select the trigger on the graph, and then choose **Remove**.
   - Choose **Add new**, and complete the **Add trigger** form.
     1. For **Trigger type**, select **Schedule**, **On demand**, or **EventBridge event**.

        For trigger type **Schedule**, choose one of the **Frequency** options. Choose **Custom** to enter a `cron` expression.

        For trigger type **EventBridge event**, enter **Number of events** (batch size), and optionally enter **Time delay** (batch window). If you omit **Time delay**, the batch window defaults to 15 minutes. For more information, see Overview of workflows in AWS Glue.
     2. Choose **Add**.

The trigger appears on the graph, along with a placeholder node (labeled **Add node**). In the example below, the start trigger is a schedule trigger named `Month-close1`.

At this point, the trigger isn't saved yet.



3. If you added a new trigger, complete these steps:

   a. Do one of the following:

      - Choose the placeholder node (**Add node**).
      - Ensure that the start trigger is selected, and on the **Action** menu above the graph, choose **Add jobs/crawlers to trigger**.

   b. In the **Add jobs(s) and crawler(s) to trigger** dialog box, select one or more jobs or crawlers, and then choose **Add**.

      The trigger is saved, and the selected jobs or crawlers appear on the graph with connectors from the trigger.

      If you mistakenly added the wrong jobs or crawlers, you can select either the trigger or a connector and choose **Remove**.

Step 3: Add more triggers

Continue to build out your workflow by adding more triggers of type **Event**. To zoom in or out or to enlarge the graph canvas, use the icons to the right of the graph. For each trigger to add, complete the following steps:

> **Note**
>
> There is no action to save the workflow. After you add your last trigger and assign actions to the trigger, the workflow is complete and saved. You can always come back later and add more nodes.

1. Do one of the following:
   - To clone an existing trigger, ensure that no node on the graph is selected, and on the **Action** menu, choose **Add trigger**.
   - To add a new trigger that watches a particular job or crawler on the graph, select the job or crawler node, and then choose the **Add trigger** placeholder node.

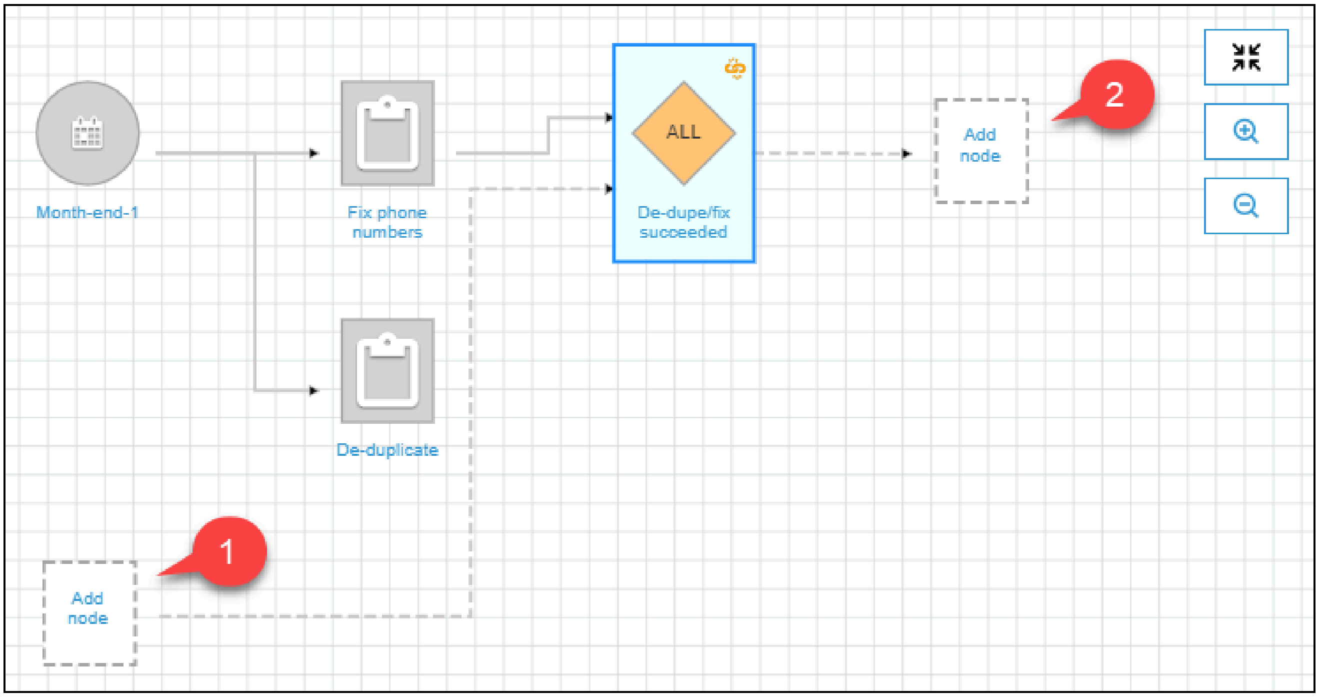
     You can add more jobs or crawlers to watch for this trigger in a later step.
2. In the **Add trigger** dialog box, do one of the following:
   - Choose **Add new**, and complete the **Add trigger** form. Then choose **Add**.

     The trigger appears on the graph. You will complete the trigger in a later step.
   - Choose **Clone existing**, and choose a trigger to clone. Then choose **Add**.

     The trigger appears on the graph, along with the jobs and crawlers that it watches and the jobs and crawlers that it starts.

     If you mistakenly chose the wrong trigger, select the trigger on the graph, and then choose **Remove**.
3. If you added a new trigger, complete these steps:
   1. Select the new trigger.

      As the following graph shows, the trigger `De-dupe/fix succeeded` is selected, and placeholder nodes appear for (1) events to watch and (2) actions.



b. (Optional if the trigger already watches an event and you want to add more jobs or crawlers to watch.) Choose the events-to-watch placeholder node, and in the **Add job(s) and crawler(s) to watch** dialog box, select one or more jobs or crawlers. Choose an event to watch (SUCCEEDED, FAILED, etc.), and choose **Add**.

c. Ensure that the trigger is selected, and choose the actions placeholder node.

d. In the **Add job(s) and crawler(s) to watch dialog** box, select one or more jobs or crawlers, and choose **Add**.

   The selected jobs and crawlers appear on the graph, with connectors from the trigger.

For more information on workflows and blueprints, see the following topics.

- Overview of workflows in AWS Glue
- Running and monitoring a workflow in AWS Glue
- Creating a workflow from a blueprint in AWS Glue

Starting an AWS Glue workflow with an Amazon EventBridge event

Amazon EventBridge, also known as CloudWatch Events, enables you to automate your AWS services and respond automatically to system events such as application availability issues or resource changes. Events from AWS services are delivered to EventBridge in near real time. You can write simple rules to indicate which events are of interest to you, and what automated actions to take when an event matches a rule.

With EventBridge support, AWS Glue can serve as an event producer and consumer in an event-driven architecture. For workflows, AWS Glue supports any type of EventBridge event as a consumer. The likely most common use case is the arrival of a new object in an Amazon S3 bucket. If you have data arriving in irregular or undefined intervals, you can process this data as close to its arrival as possible.

> **Note**
>
> AWS Glue does not provide guaranteed delivery of EventBridge messages. AWS Glue performs no deduplication if EventBridge delivers duplicate messages. You must manage idempotency based on your use case.
> Be sure to configure EventBridge rules correctly to avoid sending unwanted events.

Before you begin

If you want to start a workflow with Amazon S3 data events, you must ensure that events for the S3 bucket of interest are logged to AWS CloudTrail and EventBridge. To do so, you must create a CloudTrail trail. For more information, see Creating a trail for your AWS account.

To start a workflow with an EventBridge event

> **Note**
>
> In the following commands, replace:
>
> - *`<workflow-name>`* with the name to assign to the workflow.
> - *`<trigger-name>`* with the name to assign to the trigger.
> - *`<bucket-name>`* with the name of the Amazon S3 bucket.
> - *`<account-id>`* with a valid AWS account ID.
> - *`<region>`* with the name of the Region (for example, `us-east-1`).
> - *`<rule-name>`* with the name to assign to the EventBridge rule.

1. Ensure that you have AWS Identity and Access Management (IAM) permissions to create and view EventBridge rules and targets. The following is a sample policy that you can attach. You might want to scope it down to put limits on the operations and resources.

   JSON

   ```
   {
     "Version": "2012-10-17",
     "Statement": [
       {
         "Effect": "Allow",
         "Action": [
           "events:PutRule",
           "events:DisableRule",
           "events:DeleteRule",
           "events:PutTargets",
           "events:RemoveTargets",
           "events:EnableRule",
           "events:List*",
           "events:Describe*"
         ],
         "Resource": "*"
       }
     ]
   }
   ```

2. Create an IAM role that the EventBridge service can assume when passing an event to AWS Glue.

   a. On the **Create role** page of the IAM console, choose **AWS Service**. Then choose the service **CloudWatch Events**.

   b. Complete the **Create role** wizard. The wizard automatically attaches the `CloudWatchEventsBuiltInTargetExecutionAccess` and `CloudWatchEventsInvocationAccess` policies.

   c. Attach the following inline policy to the role. This policy allows the EventBridge service to direct events to AWS Glue.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:notifyEvent"
      ],
      "Resource": [
        "arn:aws:glue:us-east-1:111122223333:workflow/workflow-name"
      ]
    }
  ]
}
```

3. Enter the following command to create the workflow.

   See create-workflow in the *AWS CLI Command Reference* for information about additional optional command-line parameters.

   ```
   aws glue create-workflow --name <workflow-name>
   ```

4. Enter the following command to create an EventBridge event trigger for the workflow. This will be the start trigger for the workflow. Replace *`<actions>`* with the actions to perform (the jobs and crawlers to start).

   See create-trigger in the *AWS CLI Command Reference* for information about how to code the `actions` argument.

   ```
   aws glue create-trigger --workflow-name <workflow-name> --type EVENT --
   name <trigger-name> --actions <actions>
   ```

   If you want the workflow to be triggered by a batch of events instead of a single EventBridge event, enter the following command instead.

```
aws glue create-trigger --workflow-name <workflow-name> --type EVENT
 --name <trigger-name> --event-batching-condition BatchSize=<number-of-
events>,BatchWindow=<seconds> --actions <actions>
```

For the `event-batching-condition` argument, `BatchSize` is required and `BatchWindow` is optional. If `BatchWindow` is omitted, the window defaults to 900 seconds, which is the maximum window size.

Example

The following example creates a trigger that starts the `eventtest` workflow after three EventBridge events have arrived, or five minutes after the first event arrives, whichever comes first.

```
aws glue create-trigger --workflow-name eventtest --type EVENT --name objectArrival
 --event-batching-condition BatchSize=3,BatchWindow=300 --actions JobName=test1
```

5. Create a rule in Amazon EventBridge.

   a. Create the JSON object for the rule details in your preferred text editor.

      The following example specifies Amazon S3 as the event source, `PutObject` as the event name, and the bucket name as a request parameter. This rule starts a workflow when a new object arrives in the bucket.

```
{
  "source": [
    "aws.s3"
  ],
  "detail-type": [
    "AWS API Call via CloudTrail"
  ],
  "detail": {
    "eventSource": [
      "s3.amazonaws.com"
    ],
    "eventName": [
      "PutObject"
    ],
    "requestParameters": {
      "bucketName": [
        "<bucket-name>"
      ]
    }
  }
}
```

To start the workflow when a new object arrives in a folder within the bucket, you can substitute the following code for `requestParameters`.

```
    "requestParameters": {
      "bucketName": [
        "<bucket-name>"
      ]
      "key" : [{ "prefix" : "<folder1>/<folder2>/*"}}]
  }
```

b. Use your preferred tool to convert the rule JSON object to an escaped string.

```
{\n  \"source\": [\n    \"aws.s3\"\n  ],\n  \"detail-type\": [\n    \"AWS API
 Call via CloudTrail\"\n  ],\n  \"detail\": {\n    \"eventSource\": [\n
 \"s3.amazonaws.com\"\n    ],\n    \"eventName\": [\n      \"PutObject\"\n
  ],\n    \"requestParameters\": {\n      \"bucketName\": [\n        \"<bucket-
name>\"\n      ]\n    }\n  }\n}
```

c. Run the following command to create a JSON parameter template that you can edit to specify input parameters to a subsequent `put-rule` command. Save the output in a file. In this example, the file is called `ruleCommand`.

```
aws events put-rule --name <rule-name> --generate-cli-skeleton >ruleCommand
```

For more information about the `--generate-cli-skeleton` parameter, see [Generating AWS CLI skeleton and input parameters from a JSON or YAML input file]() in the *AWS Command Line Interface User Guide*.

The output file should look like the following.

```
{
    "Name": "",
    "ScheduleExpression": "",
    "EventPattern": "",
    "State": "ENABLED",
    "Description": "",
    "RoleArn": "",
    "Tags": [
        {
            "Key": "",
            "Value": ""
        }
    ],
    "EventBusName": ""
}
```

d. Edit the file to optionally remove parameters and to specify at a minimum the `Name`, `EventPattern`, and `State` parameters. For the `EventPattern` parameter, provide the escaped string for the rule details that you created in a previous step.

```
{
    "Name": "<rule-name>",
    "EventPattern": "{\n  \"source\": [\n    \"aws.s3\"\n  ],\n  \"detail-
type\": [\n    \"AWS API Call via CloudTrail\"\n  ],\n  \"detail\": {\n
 \"eventSource\": [\n      \"s3.amazonaws.com\"\n    ],\n    \"eventName\": [\n
      \"PutObject\"\n    ],\n    \"requestParameters\": {\n      \"bucketName
\": [\n        \"<bucket-name>\"\n      ]\n    }\n  }\n}",
    "State": "DISABLED",
    "Description": "Start an AWS Glue workflow upon new file arrival in an
 Amazon S3 bucket"
}
```

> **Note**
>
> It is best to leave the rule disabled until you finish building out the workflow.

e. Enter the following `put-rule` command, which reads input parameters from the file `ruleCommand`.

```
aws events put-rule --name <rule-name> --cli-input-json file://ruleCommand
```

The following output indicates success.

```
{
    "RuleArn": "<rule-arn>"
}
```

6. Enter the following command to attach the rule to a target. The target is the workflow in AWS Glue. Replace *<role-name>* with the role that you created at the beginning of this procedure.

```
aws events put-targets --rule <rule-name> --targets
 "Id"="1","Arn"="arn:aws:glue:<region>:<account-id>:workflow/<workflow-
name>","RoleArn"="arn:aws:iam::<account-id>:role/<role-name>" --region <region>
```

   The following output indicates success.

```
{
    "FailedEntryCount": 0,
    "FailedEntries": []
}
```

7. Confirm successful connection of the rule and target by entering the following command.

```
aws events list-rule-names-by-target --target-arn arn:aws:glue:<region>:<account-
id>:workflow/<workflow-name>
```

   The following output indicates success, where *<rule-name>* is the name of the rule that you created.

```
{
    "RuleNames": [
        "<rule-name>"
    ]
}
```

8. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
9. Select the workflow, and verify that the start trigger and its actions—the jobs or crawlers that it starts— appear on the workflow graph. Then continue with the procedure in Step 3: Add more triggers. Or add more components to the workflow by using the AWS Glue API or AWS Command Line Interface.
10. When the workflow is completely specified, enable the rule.

```
aws events enable-rule --name <rule-name>
```

The workflow is now ready to be started by an EventBridge event or event batch.

See also

- *Amazon EventBridge User Guide*
- Overview of workflows in AWS Glue
- Creating and building out a workflow manually in AWS Glue

Viewing the EventBridge events that started a workflow

You can view the event ID of the Amazon EventBridge event that started your workflow. If your workflow was started by a batch of events, you can view the event IDs of all events in the batch.

For workflows with a batch size greater than one, you can also see which batch condition started the workflow: the arrival of the number of events in the batch size, or the expiration of the batch window.

To view the EventBridge events that started a workflow (console)

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, choose **Workflows**.
3. Select the workflow. Then at the bottom, choose the **History** tab.
4. Select a workflow run, and then choose **View run details**.
5. On the run details page, locate the **Run properties** field, and look for the **aws:eventIds** key.

   The value for that key is a list of EventBridge event IDs.

To view the EventBridge events that started a workflow (AWS API)

- Include the following code in your Python script.

  ```
  workflow_params =
   glue_client.get_workflow_run_properties(Name=workflow_name,RunId=workflow_run_id)
  batched_events = workflow_params['aws:eventIds']
  ```

`batched_events` will be a list of strings, where each string is an event ID.

See also

- *Amazon EventBridge User Guide*
- the section called “Overview of workflows”

Running and monitoring a workflow in AWS Glue

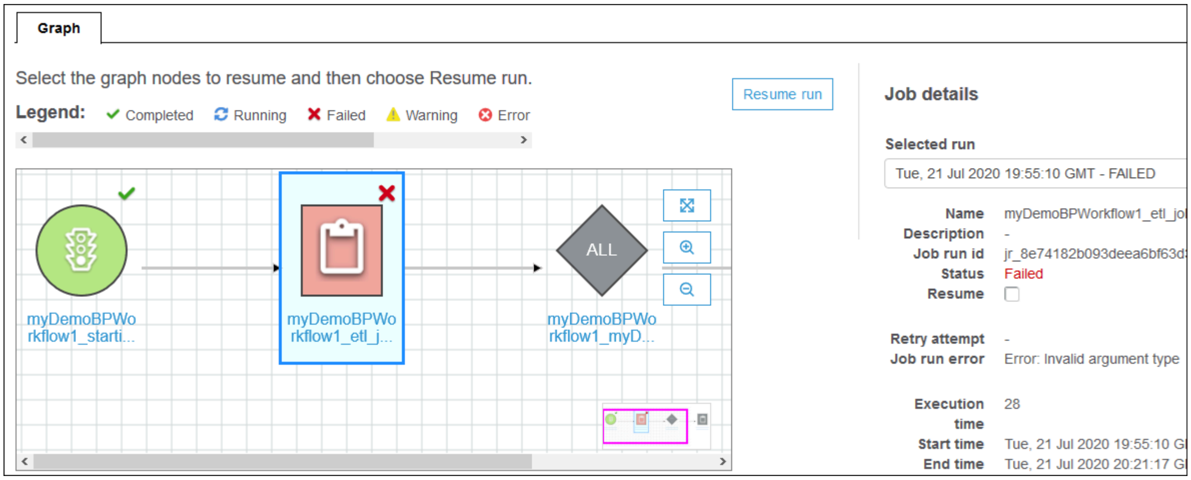
If the start trigger for a workflow is an on-demand trigger, you can start the workflow from the AWS Glue console. Complete the following steps to run and monitor a workflow. If the workflow fails, you can view the run graph to determine the node that failed. To help troubleshoot, if the workflow was created from a blueprint, you can view the blueprint run to see the blueprint parameter values that were used to create the workflow. For more information, see the section called “Viewing blueprint runs”.

You can run and monitor a workflow by using the AWS Glue console, API, or AWS Command Line Interface (AWS CLI).

To run and monitor a workflow (console)

1. Open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, under **ETL**, choose **Workflows**.
3. Select a workflow. On the **Actions** menu, choose **Run**.
4. Check the **Last run status** column in the workflows list. Choose the refresh button to view ongoing workflow status.
5. While the workflow is running or after it has completed (or failed), view the run details by completing the following steps.
   a. Ensure that the workflow is selected, and choose the **History** tab.
   b. Choose the current or most recent workflow run, and then choose **View run details**.

      The workflow runtime graph shows the current run status.
   c. Choose any node in the graph to view details and status of the node.


Graph
Legend: ✓ Completed ⟳ Running ✗ Failed ⚠ Warning ⊗ Error
Resume run
myDemoBPWorkflow1_starti...
myDemoBPWorkflow1_etl_j...
ALL
myDemoBPWorkflow1_myD...

Select the graph nodes to resume and then choose Resume run.

Job details

Selected run

Tue, 21 Jul 2020 19:55:10 GMT - FAILED

Name	myDemoBPWorkflow1_etl_jo
Description	-
Job run id	jr_8e74182b093deea6bf63d
Status	Failed
Resume	☐
Retry attempt	-
Job run error	Error: Invalid argument type
Execution time	28
Start time	Tue, 21 Jul 2020 19:55:10 G
End time	Tue, 21 Jul 2020 20:21:17 G

To run and monitor a workflow (AWS CLI)

1. Enter the following command. Replace *<workflow-name>* with the workflow to run.

   ```
   aws glue start-workflow-run --name <workflow-name>
   ```

   If the workflow is successfully started, the command returns the run ID.

2. View workflow run status by using the `get-workflow-run` command. Supply the workflow name and run ID.

   ```
   aws glue get-workflow-run --name myWorkflow --run-id
    wr_d2af14217e8eae775ba7b1fc6fc7a42c795aed3cbcd8763f9415452e2dbc8705
   ```

   The following is sample command output.

   ```
   {
       "Run": {
           "Name": "myWorkflow",
           "WorkflowRunId":
    "wr_d2af14217e8eae775ba7b1fc6fc7a42c795aed3cbcd8763f9415452e2dbc8705",
           "WorkflowRunProperties": {
               "run_state": "COMPLETED",
               "unique_id": "fee63f30-c512-4742-a9b1-7c8183bdaae2"
           },
           "StartedOn": 1578556843.049,
           "CompletedOn": 1578558649.928,
           "Status": "COMPLETED",
   ```

```
        "Statistics": {
            "TotalActions": 11,
            "TimeoutActions": 0,
            "FailedActions": 0,
            "StoppedActions": 0,
            "SucceededActions": 9,
            "RunningActions": 0,
            "ErroredActions": 0
        }
    }
}
```

See also:

- the section called “Overview of workflows”
- the section called “Overview of blueprints”

Stopping a workflow run

You can use the AWS Glue console, AWS Command Line Interface (AWS CLI) or AWS Glue API to stop a workflow run. When you stop a workflow run, all running jobs and crawlers are immediately terminated, and jobs and crawlers that are not yet started never start. It might take up to a minute for all running jobs and crawlers to stop. The workflow run status goes from **Running** to **Stopping**, and when the workflow run is completely stopped, the status goes to **Stopped**.

After the workflow run is stopped, you can view the run graph to see which jobs and crawlers completed and which never started. You can then determine if you must perform any steps to ensure data integrity. Stopping a workflow run causes no automatic rollback operations to be performed.

To stop a workflow run (console)

1. Open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, under **ETL**, choose **Workflows**.
3. Choose a running workflow, and then choose the **History** tab.
4. Choose the workflow run, and then choose **Stop run**.

The run status changes to **Stopping**.

5. (Optional) Choose the workflow run, choose **View run details**, and review the run graph.

To stop a workflow run (AWS CLI)

- Enter the following command. Replace *<workflow-name>* with the name of the workflow and *<run-id>* with the run ID of the workflow run to stop.

  ```
  aws glue stop-workflow-run --name <workflow-name> --run-id <run-id>
  ```

  The following is an example of the **stop-workflow-run** command.

  ```
  aws glue stop-workflow-run --name my-workflow --run-id
   wr_137b88917411d128081069901e4a80595d97f719282094b7f271d09576770354
  ```

Repairing and resuming a workflow run

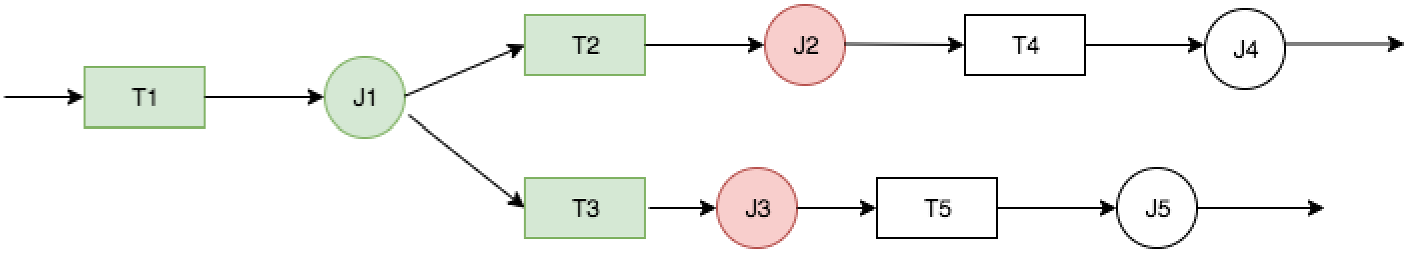
If one or more nodes (jobs or crawlers) in a workflow do not successfully complete, this means that the workflow only partially ran. After you find the root causes and make corrections, you can select one or more nodes to resume the workflow run from, and then resume the workflow run. The selected nodes and all nodes that are downstream from those nodes are then run.

Topics

- Resuming a workflow run: How it works
- Resuming a workflow run
- Notes and limitations for resuming workflow runs

Resuming a workflow run: How it works

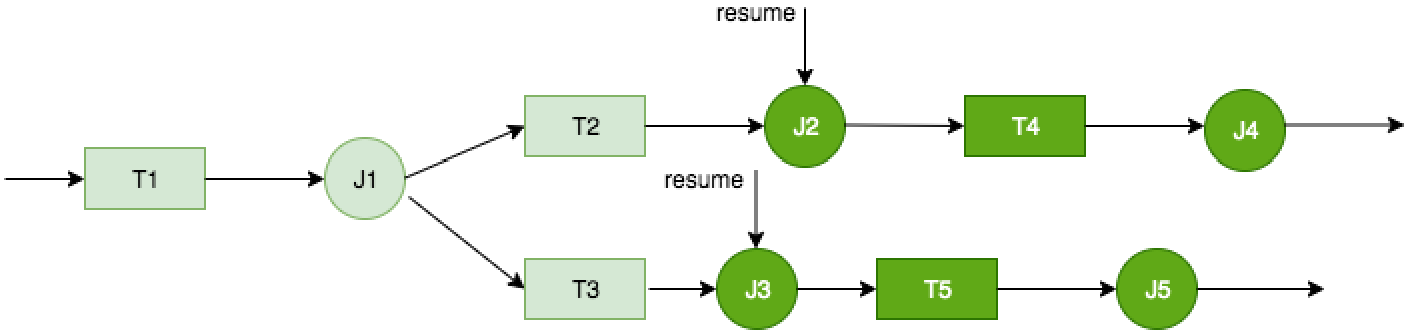
Consider the workflow W1 in the following diagram.



The workflow run proceeds as follows:

1. Trigger T1 starts job J1.
2. Successful completion of J1 fires triggers T2 and T3, which run jobs J2 and J3, respectively.
3. Jobs J2 and J3 fail.
4. Triggers T4 and T5 depend on the successful completion of J2 and J3, so they don't fire, and jobs J4 and J5 don't run. Workflow W1 is only partially run.

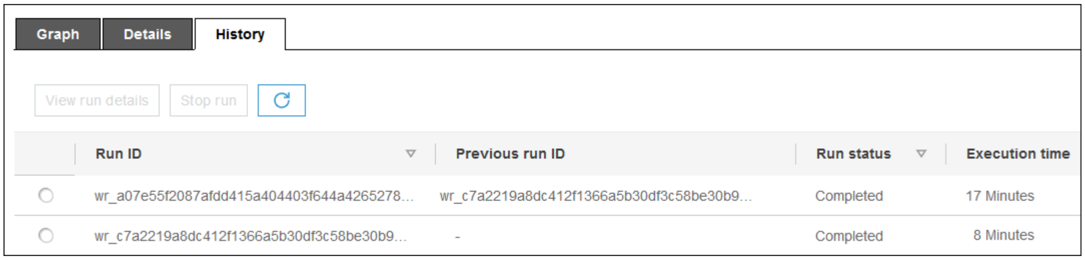
Now assume that the issues that caused J2 and J3 to fail are corrected. J2 and J3 are selected as the starting points to resume the workflow run from.



The workflow run resumes as follows:

1. Jobs J2 and J3 run successfully.
2. Triggers T4 and T5 fire.
3. Jobs J4 and J5 run successfully.

The resumed workflow run is tracked as a separate workflow run with a new run ID. When you view the workflow history, you can view the previous run ID for any workflow run. In the example in the following screenshot, the workflow run with run ID `wr_c7a22...` (the second row) had a node that did not complete. The user fixed the problem and resumed the workflow run, which resulted in run ID `wr_a07e55...` (the first row).



	Run ID	Previous run ID	Run status	Execution time
○	wr_a07e55f2087afdd415a404403f644a4265278...	wr_c7a2219a8dc412f1366a5b30df3c58be30b9...	Completed	17 Minutes
○	wr_c7a2219a8dc412f1366a5b30df3c58be30b9...	-	Completed	8 Minutes

> **Note**
>
> For the rest of this discussion, the term "resumed workflow run" refers to the workflow run that was created when the previous workflow run was resumed. The "original workflow run" refers to the workflow run that only partially ran and that needed to be resumed.

Resumed workflow run graph

In a resumed workflow run, although only a subset of nodes are run, the run graph is a complete graph. That is, the nodes that didn't run in the resumed workflow are copied from the run graph of the original workflow run. Copied job and crawler nodes that ran in the original workflow run include run details.

Consider again the workflow W1 in the previous diagram. When the workflow run is resumed starting with J2 and J3, the run graph for the resumed workflow run shows all jobs, J1 though J5, and all triggers, T1 through T5. The run details for J1 are copied from the original workflow run.

Workflow run snapshots

When a workflow run is started, AWS Glue takes a snapshot of the workflow design graph at that point in time. That snapshot is used for the duration of the workflow run. If you make changes to any triggers after the run starts, those changes don't affect the current workflow run. Snapshots ensure that workflow runs proceed in a consistent manner.

Snapshots make only triggers immutable. Changes that you make to downstream jobs and crawlers during the workflow run take effect for the current run.

Resuming a workflow run

Follow these steps to resume a workflow run. You can resume a workflow run by using the AWS Glue console, API, or AWS Command Line Interface (AWS CLI).

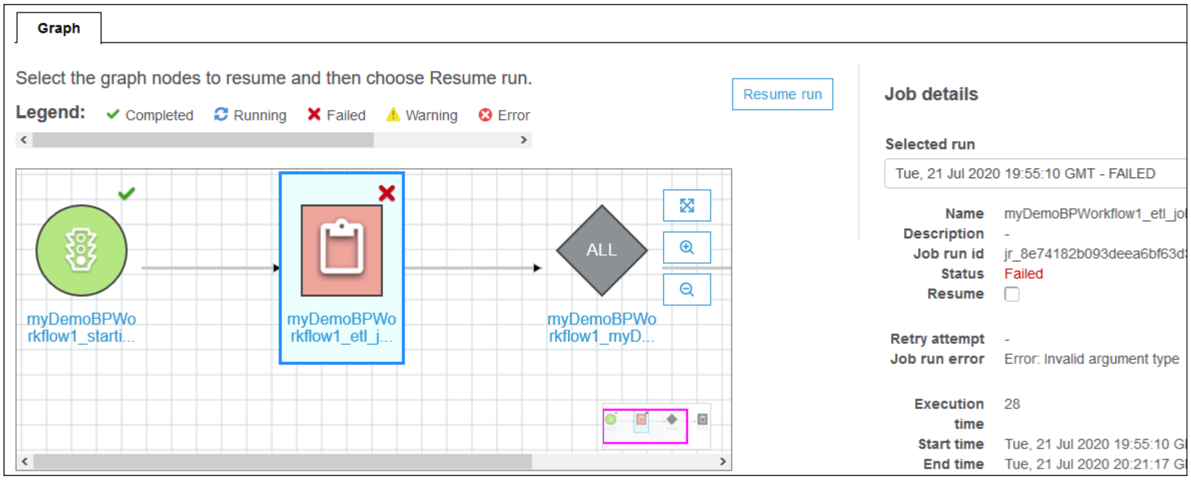
To resume a workflow run (console)

1. Open the AWS Glue console at https://console.aws.amazon.com/glue/.

   Sign in as a user who has permissions to view workflows and resume workflow runs.

   > **Note**
   >
   > To resume workflow runs, you need the `glue:ResumeWorkflowRun` AWS Identity and Access Management (IAM) permission.

2. In the navigation pane, choose **Workflows**.
3. Select a workflow, and then choose the **History** tab.
4. Select the workflow run that only partially ran, and then choose **View run details**.
5. In the run graph, select the first (or only) node that you want to restart and that you want to resume the workflow run from.
6. In the details pane to the right of the graph, select the **Resume** check box.



Select the graph nodes to resume and then choose Resume run.

Legend: ✓ Completed ↻ Running ✗ Failed ⚠ Warning ⊗ Error

Job details

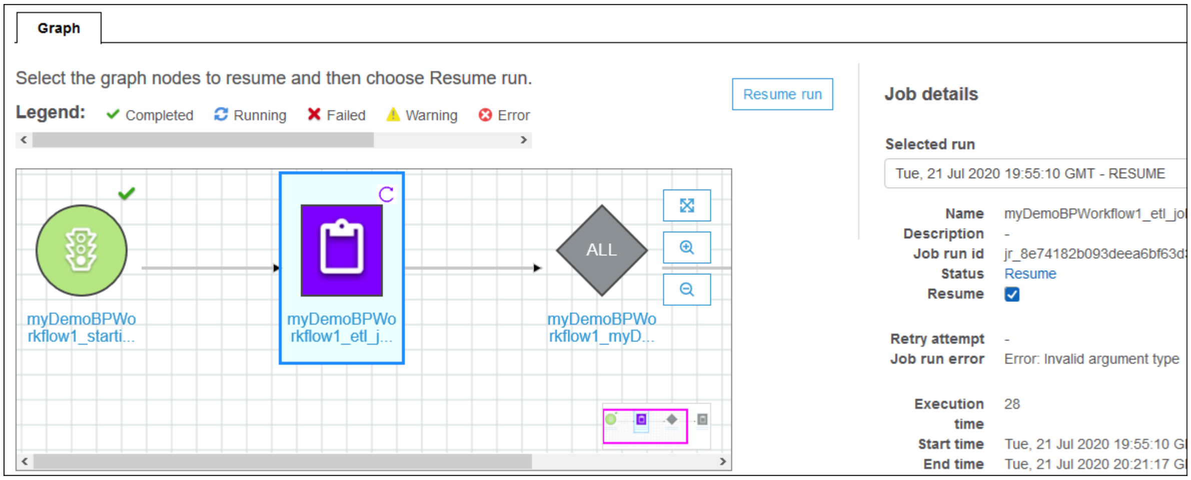
Selected run
Tue, 21 Jul 2020 19:55:10 GMT - FAILED

Name myDemoBPWorkflow1_etl_jo
Description -
Job run id jr_8e74182b093deea6bf63d
Status Failed
Resume ☐

Retry attempt -
Job run error Error: Invalid argument type

Execution time 28
Start time Tue, 21 Jul 2020 19:55:10 G
End time Tue, 21 Jul 2020 20:21:17 G

   The node changes color and shows a small resume icon at the upper right.



Select the graph nodes to resume and then choose Resume run.

Job details

Selected run

Tue, 21 Jul 2020 19:55:10 GMT - RESUME

Name myDemoBPWorkflow1_etl_jo
Description -
Job run id jr_8e74182b093deea6bf63d
Status Resume
Resume ☑

Retry attempt -
Job run error Error: Invalid argument type

Execution time 28
Start time Tue, 21 Jul 2020 19:55:10 G
End time Tue, 21 Jul 2020 20:21:17 G

7. Complete the previous two steps for any additional nodes to restart.

8. Choose **Resume run**.

To resume a workflow run (AWS CLI)

1. Ensure that you have the `glue:ResumeWorkflowRun` IAM permission.

2. Retrieve the node IDs for the nodes that you want to restart.

   a. Run the `get-workflow-run` command for the original workflow run. Supply the workflow name and run ID, and add the `--include-graph` option, as shown in the following example. Get the run ID from the **History** tab on the console, or by running the `get-workflow` command.

   ```
   aws glue get-workflow-run --name cloudtrailtest1 --run-id
    wr_a07e55f2087afdd415a404403f644a4265278f68b13ba3da08c71924ebe3c3a8 --include-
   graph
   ```

   The command returns the nodes and edges of the graph as a large JSON object.

   b. Locate the nodes of interest by the `Type` and `Name` properties of the node objects.

   The following is an example node object from the output.

   ```
   {
       "Type": "JOB",
       "Name": "test1_post_failure_4592978",
       "UniqueId":
    "wnode_d1b2563c503078b153142ee76ce545fe5ceef66e053628a786ddd74a05da86fd",
   ```

```
    "JobDetails": {
        "JobRuns": [
            {
                "Id":
 "jr_690b9f7fc5cb399204bc542c6c956f39934496a5d665a42de891e5b01f59e613",
                "Attempt": 0,
                "TriggerName": "test1_aggregate_failure_649b2432",
                "JobName": "test1_post_failure_4592978",
                "StartedOn": 1595358275.375,
                "LastModifiedOn": 1595358298.785,
                "CompletedOn": 1595358298.785,
                "JobRunState": "FAILED",
                "PredecessorRuns": [],
                "AllocatedCapacity": 0,
                "ExecutionTime": 16,
                "Timeout": 2880,
                "MaxCapacity": 0.0625,
                "LogGroupName": "/aws-glue/python-jobs"
            }
        ]
    }
}
```

   c. Get the node ID from the `UniqueId` property of the node object.

3. Run the `resume-workflow-run` command. Provide the workflow name, run ID, and list of node IDs separated by spaces, as shown in the following example.

```
aws glue resume-workflow-run --name cloudtrailtest1 --run-id
 wr_a07e55f2087afdd415a404403f644a4265278f68b13ba3da08c71924ebe3c3a8 --node-
ids wnode_ca1f63e918fb855e063aed2f42ec5762ccf71b80082ae2eb5daeb8052442f2f3
 wnode_d1b2563c503078b153142ee76ce545fe5ceef66e053628a786ddd74a05da86fd
```

The command outputs the run ID of the resumed (new) workflow run and a list of nodes that will be started.

```
{
    "RunId": "wr_2ada0d3209a262fc1156e4291134b3bd643491bcfb0ceead30bd3e4efac24de9",
    "NodeIds": [
        "wnode_ca1f63e918fb855e063aed2f42ec5762ccf71b80082ae2eb5daeb8052442f2f3"
    ]
}
```

Note that although the example `resume-workflow-run` command listed two nodes to restart, the example output indicated that only one node would be restarted. This is because one node was downstream of the other node, and the downstream node would be restarted anyway by the normal flow of the workflow.

Notes and limitations for resuming workflow runs

Keep the following notes and limitations in mind when resuming workflow runs.

- You can resume a workflow run only if it's in the `COMPLETED` state.

> **Note**
>
> Even if one ore more nodes in a workflow run don't complete, the workflow run state is shown as `COMPLETED`. Be sure to check the run graph to discover any nodes that didn't successfully complete.

- You can resume a workflow run from any job or crawler node that the original workflow run attempted to run. You can't resume a workflow run from a trigger node.
- Restarting a node does not reset its state. Any data that was partially processed is not rolled back.
- You can resume the same workflow run multiple times. If a resumed workflow run only partially runs, you can address the issue and resume the resumed run.
- If you select two nodes to restart and they're dependent upon each other, the upstream node is run before the downstream node. In fact, selecting the downstream node is redundant, because it will be run according to the normal flow of the workflow.

Getting and setting workflow run properties in AWS Glue

Use workflow run properties to share and manage state among the jobs in your AWS Glue workflow. You can set default run properties when you create the workflow. Then, as your jobs run, they can retrieve the run property values and optionally modify them for input to jobs that are later in the workflow. When a job modifies a run property, the new value exists only for the workflow run. The default run properties aren't affected.

If your AWS Glue job is not part of a workflow, these properties will not be set.

The following sample Python code from an extract, transform, and load (ETL) job demonstrates how to get the workflow run properties.

```
import sys
import boto3
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from awsglue.context import GlueContext
from pyspark.context import SparkContext

glue_client = boto3.client("glue")
args = getResolvedOptions(sys.argv, ['JOB_NAME','WORKFLOW_NAME', 'WORKFLOW_RUN_ID'])
workflow_name = args['WORKFLOW_NAME']
workflow_run_id = args['WORKFLOW_RUN_ID']
workflow_params = glue_client.get_workflow_run_properties(Name=workflow_name,
                                        RunId=workflow_run_id)["RunProperties"]

target_database = workflow_params['target_database']
target_s3_location = workflow_params['target_s3_location']
```

The following code continues by setting the `target_format` run property to `'csv'`.

```
workflow_params['target_format'] = 'csv'
glue_client.put_workflow_run_properties(Name=workflow_name, RunId=workflow_run_id,
 RunProperties=workflow_params)
```

For more information, see the following:

- GetWorkflowRunProperties action (Python: get_workflow_run_properties)
- PutWorkflowRunProperties action (Python: put_workflow_run_properties)

Querying workflows using the AWS Glue API

AWS Glue provides a rich API for managing workflows. You can retrieve a static view of a workflow or a dynamic view of a running workflow using the AWS Glue API. For more information, see Workflows.

Topics

- Querying static views
- Querying dynamic views

Querying static views

Use the `GetWorkflow` API operation to get a static view that indicates the design of a workflow. This operation returns a directed graph consisting of nodes and edges, where a node represents a trigger, a job, or a crawler. Edges define the relationships between nodes. They are represented by connectors (arrows) on the graph in the AWS Glue console.

You can also use this operation with popular graph-processing libraries such as NetworkX, igraph, JGraphT, and the Java Universal Network/Graph (JUNG) Framework. Because all these libraries represent graphs similarly, minimal transformations are needed.

The static view returned by this API is the most up-to-date view according to the latest definition of triggers associated with the workflow.

Graph definition

A workflow graph G is an ordered pair (N, E), where N is a set of nodes and E a set of edges. *Node* is a vertex in the graph identified by a unique number. A node can be of type trigger, job, or crawler. For example: `{name:T1, type:Trigger, uniqueId:1}, {name:J1, type:Job, uniqueId:2}`.

Edge is a 2-tuple of the form (`src, dest`), where `src` and `dest` are nodes and there is a directed edge from `src` to `dest`.

Example of querying a static view

Consider a conditional trigger T, which triggers job J2 upon completion of job J1.

```
J1 ---> T ---> J2
```

Nodes: J1, T, J2

Edges: (J1, T), (T, J2)

Querying dynamic views

Use the `GetWorkflowRun` API operation to get a dynamic view of a running workflow. This operation returns the same static view of the graph along with metadata related to the workflow run.

For run, nodes representing jobs in the `GetWorkflowRun` call have a list of job runs initiated as part of the latest run of the workflow. You can use this list to display the run status of each job in the graph itself. For downstream dependencies that are not yet run, this field is set to `null`. The graphed information makes you aware of the current state of any workflow at any point of time.

The dynamic view returned by this API is based on the static view that was present when the workflow run was started.

Runtime nodes example: `{name:T1, type: Trigger, uniqueId:1}`, `{name:J1, type:Job, uniqueId:2, jobDetails:{jobRuns}}`, `{name:C1, type:Crawler, uniqueId:3, crawlerDetails:{crawls}}`

Example 1: Dynamic view

The following example illustrates a simple two-trigger workflow.

- Nodes: t1, j1, t2, j2
- Edges: (t1, j1), (j1, t2), (t2, j2)

The `GetWorkflow` response contains the following.

```
{
    Nodes : [
        {
            "type" : Trigger,
            "name" : "t1",
            "uniqueId" : 1
        },
        {
            "type" : Job,
            "name" : "j1",
            "uniqueId" : 2
        },
        {
            "type" : Trigger,
            "name" : "t2",
            "uniqueId" : 3
        },
        {
            "type" : Job,
            "name" : "j2",
            "uniqueId" : 4
        }
    ],
    Edges : [
        {
            "sourceId" : 1,
            "destinationId" : 2
        },
        {
            "sourceId" : 2,
            "destinationId" : 3
        },
        {
            "sourceId" : 3,
            "destinationId" : 4
        }
}
```

The `GetWorkflowRun` response contains the following.

```
{
    Nodes : [
        {
            "type" : Trigger,
            "name" : "t1",
            "uniqueId" : 1,
            "jobDetails" : null,
            "crawlerDetails" : null
        },
        {
            "type" : Job,
            "name" : "j1",
            "uniqueId" : 2,
            "jobDetails" : [
                {
                    "id" : "jr_12334",
                    "jobRunState" : "SUCCEEDED",
                    "errorMessage" : "error string"
                }
            ],
            "crawlerDetails" : null
        },
        {
            "type" : Trigger,
            "name" : "t2",
            "uniqueId" : 3,
            "jobDetails" : null,
            "crawlerDetails" : null
        },
        {
            "type" : Job,
            "name" : "j2",
            "uniqueId" : 4,
            "jobDetails" : [
                {
                    "id" : "jr_1233sdf4",
                    "jobRunState" : "SUCCEEDED",
                    "errorMessage" : "error string"
                }
            ],
            "crawlerDetails" : null
        }
    ],
    Edges : [
        {
            "sourceId" : 1,
            "destinationId" : 2
        },
        {
            "sourceId" : 2,
            "destinationId" : 3
        },
        {
            "sourceId" : 3,
            "destinationId" : 4
        }
}
```

Example 2: Multiple jobs with a conditional trigger

The following example shows a workflow with multiple jobs and a conditional trigger (t3).

```
Consider Flow:
T(t1) ---> J(j1) ---> T(t2) ---> J(j2)
             |                     |
             |                     |
             >+------> T(t3) <-----+
                          |
                          |
                        J(j3)

Graph generated:
Nodes: t1, t2, t3, j1, j2, j3
Edges: (t1, j1), (j1, t2), (t2, j2), (j1, t3), (j2, t3), (t3, j3)
```

Blueprint and workflow restrictions in AWS Glue

The following are restrictions for blueprints and workflows.

Blueprint restrictions

Keep the following blueprint restrictions in mind:

- The blueprint must be registered in the same AWS Region where the Amazon S3 bucket resides in.
- To share blueprints across AWS accounts you must give the read permissions on the blueprint ZIP archive in Amazon S3. Customers who have read permission on a blueprint ZIP archive can register the blueprint in their AWS account and use it.
- The set of blueprint parameters is stored as a single JSON object. The maximum length of this object is 128 KB.
- The maximum uncompressed size of the blueprint ZIP archive is 5 MB. The maximum compressed size is 1 MB.
- Limit the total number of jobs, crawlers, and triggers within a workflow to 100 or less. If you include more than 100, you might get errors when trying to resume or stop workflow runs.

Workflow restrictions

Keep the following workflow restrictions in mind. Some of these comments are directed more at a user creating workflows manually.

- The maximum batch size for an Amazon EventBridge event trigger is 100. The maximum window size is 900 seconds (15 minutes).
- A trigger can be associated with only one workflow.
- Only one starting trigger (on-demand or schedule) is permitted.

- If a job or crawler in a workflow is started by a trigger that is outside the workflow, any triggers inside the workflow that depend on job or crawler completion (succeeded or otherwise) do not fire.
- Similarly, if a job or crawler in a workflow has triggers that depend on job or crawler completion (succeeded or otherwise) both within the workflow and outside the workflow, and if the job or crawler is started from within a workflow, only the triggers inside the workflow fire upon job or crawler completion.

Troubleshooting blueprint errors in AWS Glue

If you encounter errors when using AWS Glue blueprints, use the following solutions to help you find the source of the problems and fix them.

Topics

- Error: missing PySpark module
- Error: missing blueprint config file
- Error: missing imported file
- Error: not authorized to perform iamPassRole on resource
- Error: invalid cron schedule
- Error: a trigger with the same name already exists
- Error: workflow with name: foo already exists.
- Error: module not found in specified layoutGenerator path
- Error: validation error in Connections field

Error: missing PySpark module

AWS Glue returns the error "Unknown error executing layout generator function ModuleNotFoundError: No module named 'pyspark'".

When you unzip the blueprint archive it could be like either of the following:

```
$ unzip compaction.zip
Archive:  compaction.zip
   creating: compaction/
  inflating: compaction/blueprint.cfg
  inflating: compaction/layout.py
  inflating: compaction/README.md
  inflating: compaction/compaction.py

$ unzip compaction.zip
Archive:  compaction.zip
  inflating: blueprint.cfg
  inflating: compaction.py
  inflating: layout.py
  inflating: README.md
```

In the first case, all the files related to the blueprint were placed under a folder named compaction and it was then converted into a zip file named *compaction.zip*.

In the second case, all the files required for the blueprint were not included into a folder and were added as root files under the zip file *compaction.zip*.

Creating a file in either of the above formats is allowed. However make sure that `blueprint.cfg` has the correct path to the name of the function in the script that generates the layout.

Examples

In case 1: `blueprint.cfg` should have `layoutGenerator` as the following:

```
layoutGenerator": "compaction.layout.generate_layout"
```

In case 2: `blueprint.cfg` should have `layoutGenerator` as the following

```
layoutGenerator": "layout.generate_layout"
```

If this path is not included correctly, you could see an error as indicated. For example, if you have the folder structure as mentioned in case 2 and you have the `layoutGenerator` indicated as in case 1, you can see the above error.

Error: missing blueprint config file

AWS Glue returns the error "Unknown error executing layout generator function FileNotFoundError: [Errno 2] No such file or directory: '/tmp/compaction/blueprint.cfg'".

The blueprint.cfg should be placed at the root level of the ZIP archive or within a folder which has the same name as the ZIP archive.

When we extract the blueprint ZIP archive, blueprint.cfg is expected to be found in one of the following paths. If it is not found in one of the following paths, you can see the above error.

```
$ unzip compaction.zip
Archive:  compaction.zip
   creating: compaction/
  inflating: compaction/blueprint.cfg

$ unzip compaction.zip
Archive:  compaction.zip
  inflating: blueprint.cfg
```

Error: missing imported file

AWS Glue returns the error "Unknown error executing layout generator function FileNotFoundError: [Errno 2] No such file or directory:* *'demo-project/foo.py'".

If your layout generation script has functionality to read other files, make sure you give a full path for the file to be imported. For example, the Conversion.py script may be referenced in Layout.py. For more information, see Sample blueprint Project.

Error: not authorized to perform iamPassRole on resource

AWS Glue returns the error "User: arn:aws:sts::123456789012:assumed-role/AWSGlueServiceRole/GlueSession is not authorized to perform: iam:PassRole on resource: arn:aws:iam::123456789012:role/AWSGlueServiceRole"

If the jobs and crawlers in the workflow assume the same role as the role passed to create workflow from the blueprint, then the blueprint role needs to include the `iam:PassRole` permission on itself.

If the jobs and crawlers in the workflow assume a role other than the role passed to create the entities of the workflow from the blueprint, then the blueprint role needs to include the `iam:PassRole` permission on that other role instead of on the blueprint role.

For more information, see Permissions for blueprint Roles.

Error: invalid cron schedule

AWS Glue returns the error "The schedule cron(0 0 * * * *) is invalid."

Provide a valid cron expression. For more information, see Time-Based Schedules for Jobs and Crawlers.

Error: a trigger with the same name already exists

AWS Glue returns the error "Trigger with name 'foo_starting_trigger' already submitted with different configuration".

A blueprint does not require you to define triggers in the layout script for workflow creation. Trigger creation is managed by the blueprint library based on the dependencies defined between two actions.

The naming for the triggers is as follows:

- For the starting trigger in the workflow the naming is <workflow_name>_starting_trigger.
- For a node(job/crawler) in the workflow that depends on the completion of either one or multiple upstream nodes; AWS Glue defines a trigger with the name <workflow_name>_<node_name>_trigger

This error means a trigger with same name already exists. You can delete the existing trigger and re-run the workflow creation.

> **Note**
>
> Deleting a workflow doesn’t delete the nodes within the workflow. It is possible that though the workflow is deleted, triggers are left behind. Due to this, you may not receive a 'workflow already exists' error, but you may receive a 'trigger already exists' error in a case where you create a workflow, delete it and then try to re-create it with the same name from same blueprint.

Error: workflow with name: foo already exists.

The workflow name should be unique. Please try with a different name.

Error: module not found in specified layoutGenerator path

AWS Glue returns the error "Unknown error executing layout generator function ModuleNotFoundError: No module named 'crawl_s3_locations'".

```
layoutGenerator": "crawl_s3_locations.layout.generate_layout"
```

For example, if you have the above layoutGenerator path, then when you unzip the blueprint archive, it needs to look like the following:

```
$ unzip crawl_s3_locations.zip
Archive:  crawl_s3_locations.zip
   creating: crawl_s3_locations/
  inflating: crawl_s3_locations/blueprint.cfg
  inflating: crawl_s3_locations/layout.py
  inflating: crawl_s3_locations/README.md
```

When you unzip the archive, if the blueprint archive looks like the following, then you can get the above error.

```
$ unzip crawl_s3_locations.zip
Archive:  crawl_s3_locations.zip
  inflating: blueprint.cfg
  inflating: layout.py
  inflating: README.md
```

You can see that there is no folder named `crawl_s3_locations` and when the `layoutGenerator` path refers to the layout file via the module `crawl_s3_locations`, you can get the above error.

Error: validation error in Connections field

AWS Glue returns the error "Unknown error executing layout generator function TypeError: Value ['foo'] for key Connections should be of type <class 'dict'>!".

This is a validation error. The `Connections` field in the `Job` class is expecting a dictionary and instead a list of values are provided causing the error.

```
User input was list of values
Connections= ['string']

Should be a dict like the following
Connections*=*{'Connections': ['string']}
```

To avoid these run time errors while creating a workflow from a blueprint, you can validate the workflow, job and crawler definitions as outlined in Testing a blueprint.

Refer to the syntax in AWS Glue blueprint Classes Reference for defining the AWS Glue job, crawler and workflow in the layout script.

Permissions for personas and roles for AWS Glue blueprints

The following are the typical personas and suggested AWS Identity and Access Management (IAM) permissions policies for personas and roles for AWS Glue blueprints.

Topics

- Blueprint personas
- Permissions for blueprint personas
- Permissions for blueprint roles

Blueprint personas

The following are the personas typically involved in the lifecycle of AWS Glue blueprints.

Persona	Description
AWS Glue developer	Develops, tests, and publishes blueprints.
AWS Glue administrator	Registers, maintains, and grants permissions on blueprints.
Data analyst	Runs blueprints to create workflows.

For more information, see the section called “Overview of blueprints”.

Permissions for blueprint personas

The following are the suggested permissions for each blueprint persona.

AWS Glue developer permissions for blueprints

The AWS Glue developer must have write permissions on the Amazon S3 bucket that is used to publish the blueprint. Often, the developer registers the blueprint after uploading it. In that case, the developer needs the permissions listed in the section called “AWS Glue administrator permissions for blueprints”. Additionally, if the developer wishes to test the blueprint after its registered, he or she also needs the permissions listed in the section called “Data analyst permissions for blueprints”.

AWS Glue administrator permissions for blueprints

The following policy grants permissions to register, view, and maintain AWS Glue blueprints.

> **Important**
>
> In the following policy, replace *<s3-bucket-name>* and *<prefix>* with the Amazon S3 path to uploaded blueprint ZIP archives to register.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:CreateBlueprint",
        "glue:UpdateBlueprint",
        "glue:DeleteBlueprint",
        "glue:GetBlueprint",
        "glue:ListBlueprints",
        "glue:BatchGetBlueprints"
      ],
      "Resource": "*"
    },
    {
      "Effect": "Allow",
      "Action": [
        "s3:GetObject"
      ],
      "Resource": "arn:aws:s3:::amzn-s3-demo-bucket/prefix/*"
    }
  ]
}
```

Data analyst permissions for blueprints

The following policy grants permissions to run blueprints and to view the resulting workflow and workflow components. It also grants `PassRole` on the role that AWS Glue assumes to create the workflow and workflow components.

The policy grants permissions on any resource. If you want to configure fine-grained access to individual blueprints, use the following format for blueprint ARNs:

```
arn:aws:glue:<region>:<account-id>:blueprint/<blueprint-name>
```

Important

In the following policy, replace *<account-id>* with a valid AWS account and replace *<role-name>* with the name of the role used to run a blueprint. See the section called "Permissions for blueprint roles" for the permissions that this role requires.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListBlueprints",
        "glue:GetBlueprint",
        "glue:StartBlueprintRun",
        "glue:GetBlueprintRun",
        "glue:GetBlueprintRuns",
        "glue:GetCrawler",
        "glue:ListTriggers",
        "glue:ListJobs",
        "glue:BatchGetCrawlers",
        "glue:GetTrigger",
        "glue:BatchGetWorkflows",
        "glue:BatchGetTriggers",
        "glue:BatchGetJobs",
        "glue:BatchGetBlueprints",
        "glue:GetWorkflowRun",
        "glue:GetWorkflowRuns",
        "glue:ListCrawlers",
        "glue:ListWorkflows",
        "glue:GetJob",
        "glue:GetWorkflow",
        "glue:StartWorkflowRun"
      ],
      "Resource": "*"
    },
    {
      "Effect": "Allow",
      "Action": "iam:PassRole",
      "Resource": "arn:aws:iam::111122223333:role/role-name"
    }
  ]
}
```

Permissions for blueprint roles

The following are the suggested permissions for the IAM role used to create a workflow from a blueprint. The role has to have a trust relationship with `glue.amazonaws.com`.

> ⚠ **Important**
>
> In the following policy, replace *<account-id>* with a valid AWS account, and replace *<role-name>* with the name of the role.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:CreateJob",
        "glue:GetCrawler",
        "glue:GetTrigger",
        "glue:DeleteCrawler",
            "glue:CreateTrigger",
            "glue:DeleteTrigger",
            "glue:DeleteJob",
            "glue:CreateWorkflow",
            "glue:DeleteWorkflow",
            "glue:GetJob",
            "glue:GetWorkflow",
            "glue:CreateCrawler"
          ],
          "Resource": "*"
        },
        {
          "Effect": "Allow",
          "Action": "iam:PassRole",
          "Resource": "arn:aws:iam::111122223333:role/role-name"
        }
      ]
    }
```

Note

If the jobs and crawlers in the workflow assume a role other than this role, this policy must include the `iam:PassRole` permission on that other role instead of on the blueprint role.

Developing blueprints in AWS Glue

Your organization might have a set of similar ETL use cases that could benefit from being able to parametrize a single workflow to handle them all. To address this need, AWS Glue enables you to define *blueprints*, which you can use to generate workflows. A blueprint accepts parameters, so that from a single blueprint, a data analyst can create different workflows to handle similar ETL use cases. After you create a blueprint, you can reuse it for different departments, teams, and projects.

Topics

- Overview of blueprints in AWS Glue
- Developing blueprints in AWS Glue
- Registering a blueprint in AWS Glue
- Viewing blueprints in AWS Glue

- Updating a blueprint in AWS Glue
- Creating a workflow from a blueprint in AWS Glue
- Viewing blueprint runs in AWS Glue

Overview of blueprints in AWS Glue

> **Note**
>
> The blueprints feature is currently unavailable in the following Regions in the AWS Glue console: Asia Pacific (Jakarta) and Middle East (UAE).

AWS Glue blueprints provide a way to create and share AWS Glue workflows. When there is a complex ETL process that could be used for similar use cases, rather than creating an AWS Glue workflow for each use case, you can create a single blueprint.

The blueprint specifies the jobs and crawlers to include in a workflow, and specifies parameters that the workflow user supplies when they run the blueprint to create a workflow. The use of parameters enables a single blueprint to generate workflows for the various similar use cases. For more information about workflows, see Overview of workflows in AWS Glue.

The following are example use cases for blueprints:

- You want to partition an existing dataset. The input parameters to the blueprint are Amazon Simple Storage Service (Amazon S3) source and target paths and a list of partition columns.
- You want to snapshot an Amazon DynamoDB table into a SQL data store like Amazon Redshift. The input parameters to the blueprint are the DynamoDB table name and an AWS Glue connection, which designates an Amazon Redshift cluster and destination database.
- You want to convert CSV data in multiple Amazon S3 paths to Parquet. You want the AWS Glue workflow to include a separate crawler and job for each path. The input parameters are the destination database in the AWS Glue Data Catalog and a comma-delimited list of Amazon S3 paths. Note that in this case, the number of crawlers and jobs that the workflow creates is variable.

Blueprint components

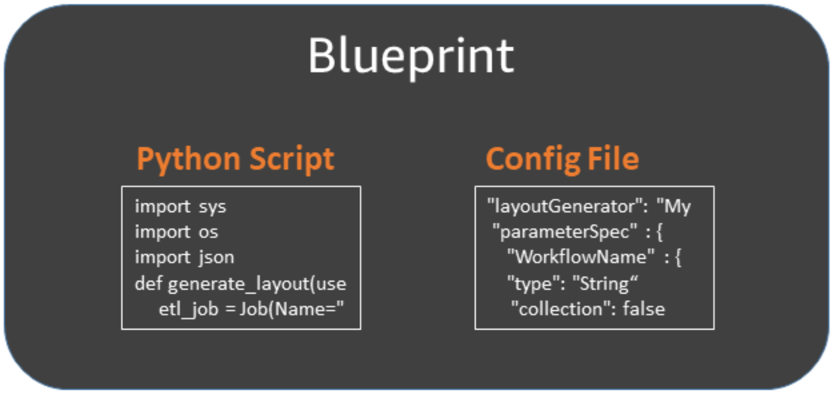
A blueprint is a ZIP archive that contains the following components:

- A Python layout generator script

  Contains a function that specifies the workflow *layout*—the crawlers and jobs to create for the workflow, the job and crawler properties, and the dependencies between the jobs and crawlers. The function accepts blueprint parameters and returns a workflow structure (JSON object) that AWS Glue uses to generate the workflow. Because you use a Python script to generate the workflow, you can add your own logic that is suitable for your use cases.
- A configuration file

  Specifies the fully qualified name of the Python function that generates the workflow layout. Also specifies the names, data types, and other properties of all blueprint parameters used by the script.
- (Optional) ETL scripts and supporting files

  As an advanced use case, you can parameterize the location of the ETL scripts that your jobs use. You can include job script files in the ZIP archive and specify a blueprint parameter for an Amazon S3 location where the scripts are to be copied to. The layout generator script can copy the ETL scripts to the designated location and specify that location as the job script location property. You can also include any libraries or other supporting files, provided that your script handles them.

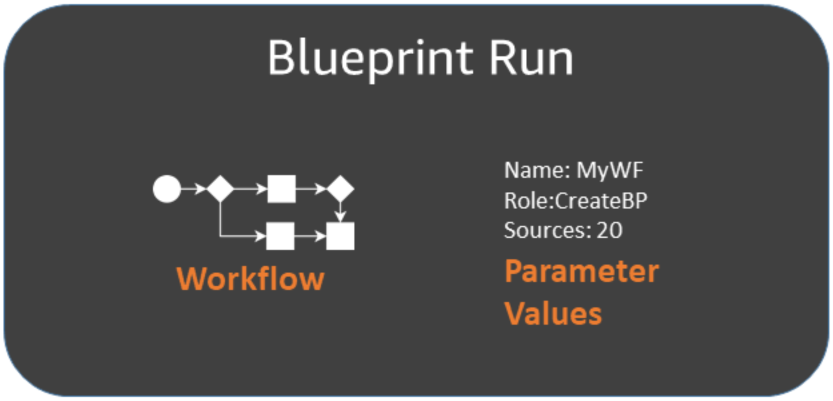


```
import sys
import os
import json
def generate_layout(use
    etl_job = Job(Name="
```

```
"layoutGenerator": "My
 "parameterSpec" : {
   "WorkflowName" : {
   "type": "String“
    "collection": false
```

Blueprint runs

When you create a workflow from a blueprint, AWS Glue runs the blueprint, which starts an asynchronous process to create the workflow and the jobs, crawlers, and triggers that the workflow encapsulates. AWS Glue uses the blueprint run to orchestrate the creation of the workflow and its components. You view the status of the creation process by viewing the blueprint run status. The blueprint run also stores the values that you supplied for the blueprint parameters.



You can view blueprint runs using the AWS Glue console or AWS Command Line Interface (AWS CLI). When viewing or troubleshooting a workflow, you can always return to the blueprint run to view the blueprint parameter values that were used to create the workflow.

Lifecycle of a blueprint

blueprints are developed, tested, registered with AWS Glue, and run to create workflows. There are typically three personas involved in the blueprint lifecycle.

Persona	Tasks
AWS Glue developer	• Writes the workflow layout script and creates the configuration file. • Tests the blueprint locally using libraries provided by the AWS Glue service. • Creates a ZIP archive of the script, configuration file, and supporting files and publishes the archive to a location in Amazon S3. • Adds a bucket policy to the Amazon S3 bucket that grants read permissions on bucket objects to the AWS Glue administrator's AWS account. • Grants IAM read permissions on the ZIP archive in Amazon S3 to the AWS Glue administrator.
AWS Glue administrator	• *Registers* the blueprint with AWS Glue. AWS Glue makes a copy of the ZIP archive into a reserved Amazon S3 location.

Persona	Tasks
	• Grants IAM permissions on the blueprint to data analysts.
Data analyst	• Runs the blueprint to create a workflow, and provides blueprint parameter values. Checks the blueprint run status to ensure that the workflow and workflow components were successfully generated. • Runs and troubleshoots the workflow. Before running the workflow, can verify the workflow by viewing the workflow design graph on the AWS Glue console.

> **See also**
>
> - Developing blueprints in AWS Glue
> - Creating a workflow from a blueprint in AWS Glue
> - Permissions for personas and roles for AWS Glue blueprints

Developing blueprints in AWS Glue

As an AWS Glue developer, you can create and publish blueprints that data analysts can use to generate workflows.

Topics

- Overview of developing blueprints
- Prerequisites for developing blueprints
- Writing the blueprint code
- Sample blueprint project
- Testing a blueprint
- Publishing a blueprint
- AWS Glue blueprint classes reference
- Blueprint samples

See also

- Overview of blueprints in AWS Glue

Overview of developing blueprints

The first step in your development process is to identify a common use case that would benefit from a blueprint. A typical use case involves a recurring ETL problem that you believe should be solved in a general manner. Next, design a blueprint that implements the generalized use case, and define the blueprint input parameters that together can define a specific use case from the generalized use case.

A blueprint consists of a project that contains a blueprint parameter configuration file and a script that defines the *layout* of the workflow to generate. The layout defines the jobs and crawlers (or *entities* in blueprint script terminology) to create.

You do not directly specify any triggers in the layout script. Instead you write code to specify the dependencies between the jobs and crawlers that the script creates. AWS Glue generates the triggers based on your dependency specifications. The output of the layout script is a workflow object, which contains specifications for all workflow entities.

You build your workflow object using the following AWS Glue blueprint libraries:

- `awsglue.blueprint.base_resource` – A library of base resources used by the libraries.
- `awsglue.blueprint.workflow` – A library for defining a `Workflow` class.
- `awsglue.blueprint.job` – A library for defining a `Job` class.
- `awsglue.blueprint.crawler` – A library for defining a `Crawler` class.

The only other libraries that are supported for layout generation are those libraries that are available for the Python shell.

Before publishing your blueprint, you can use methods defined in the blueprint libraries to test the blueprint locally.

When you're ready to make the blueprint available to data analysts, you package the script, the parameter configuration file, and any supporting files, such as additional scripts and libraries, into a single deployable asset. You then upload the asset to Amazon S3 and ask an administrator to register it with AWS Glue.

For information about more sample blueprint projects, see Sample blueprint project and Blueprint samples.

Prerequisites for developing blueprints

To develop blueprints, you should be familiar with using AWS Glue and writing scripts for Apache Spark ETL jobs or Python shell jobs. In addition, you must complete the following setup tasks.

- Download four AWS Python libraries to use in your blueprint layout scripts.
- Set up the AWS SDKs.
- Set up the AWS CLI.

Download the Python libraries

Download the following libraries from GitHub, and install them into your project:

- https://github.com/awslabs/aws-glue-blueprint-libs/tree/master/awsglue/blueprint/base_resource.py
- https://github.com/awslabs/aws-glue-blueprint-libs/tree/master/awsglue/blueprint/workflow.py
- https://github.com/awslabs/aws-glue-blueprint-libs/tree/master/awsglue/blueprint/crawler.py
- https://github.com/awslabs/aws-glue-blueprint-libs/tree/master/awsglue/blueprint/job.py

Set up the AWS Java SDK

For the AWS Java SDK, you must add a `jar` file that includes the API for blueprints.

1. If you haven't already done so, set up the AWS SDK for Java.
    - For Java 1.x, follow the instructions in Set up the AWS SDK for Java in the *AWS SDK for Java Developer Guide*.
    - For Java 2.x, follow the instructions in Setting up the AWS SDK for Java 2.x in the *AWS SDK for Java 2.x Developer Guide*.
2. Download the client `jar` file that has access to the APIs for blueprints.

  - For Java 1.x: s3://awsglue-custom-blueprints-preview-artifacts/awsglue-java-sdk-preview/AWSGlueJavaClient-1.11.x.jar
  - For Java 2.x: s3://awsglue-custom-blueprints-preview-artifacts/awsglue-java-sdk-v2-preview/AwsJavaSdk-Glue-2.0.jar

3. Add the client `jar` to the front of the Java classpath to override the AWS Glue client provided by the AWS Java SDK.

   ```
   export CLASSPATH=<path-to-preview-client-jar>:$CLASSPATH
   ```

4. (Optional) Test the SDK with the following Java application. The application should output an empty list.

   Replace `accessKey` and `secretKey` with your credentials, and replace `us-east-1` with your Region.

   ```
   import com.amazonaws.auth.AWSCredentials;
   import com.amazonaws.auth.AWSCredentialsProvider;
   import com.amazonaws.auth.AWSStaticCredentialsProvider;
   import com.amazonaws.auth.BasicAWSCredentials;
   import com.amazonaws.services.glue.AWSGlue;
   import com.amazonaws.services.glue.AWSGlueClientBuilder;
   import com.amazonaws.services.glue.model.ListBlueprintsRequest;

   public class App{
       public static void main(String[] args) {
           AWSCredentials credentials = new BasicAWSCredentials("accessKey",
    "secretKey");
           AWSCredentialsProvider provider = new
    AWSStaticCredentialsProvider(credentials);
           AWSGlue glue = AWSGlueClientBuilder.standard().withCredentials(provider)
                   .withRegion("us-east-1").build();
           ListBlueprintsRequest request = new
    ListBlueprintsRequest().withMaxResults(2);
           System.out.println(glue.listBlueprints(request));
       }
   }
   ```

Set up the AWS Python SDK

The following steps assume that you have Python version 2.7 or later, or version 3.9 or later installed on your computer.

1. Download the following boto3 wheel file. If prompted to open or save, save the file. s3://awsglue-custom-blueprints-preview-artifacts/aws-python-sdk-preview/boto3-1.17.31-py2.py3-none-any.whl
2. Download the following botocore wheel file: s3://awsglue-custom-blueprints-preview-artifacts/aws-python-sdk-preview/botocore-1.20.31-py2.py3-none-any.whl
3. Check your Python version.

   ```
   python --version
   ```

4. Depending on your Python version, enter the following commands (for Linux):

   - For Python 2.7 or later.

     ```
     python3 -m pip install --user virtualenv
     source env/bin/activate
     ```

   - For Python 3.9 or later.

     ```
     python3 -m venv python-sdk-test
     source python-sdk-test/bin/activate
     ```

5. Install the botocore wheel file.

   ```
   python3 -m pip install <download-directory>/botocore-1.20.31-py2.py3-none-any.whl
   ```

6. Install the boto3 wheel file.

   ```
   python3 -m pip install <download-directory>/boto3-1.17.31-py2.py3-none-any.whl
   ```

7. Configure your credentials and default region in the `~/.aws/credentials` and `~/.aws/config` files. For more information, see [Configuring the AWS CLI]() in the *AWS Command Line Interface User Guide*.
8. (Optional) Test your setup. The following commands should return an empty list.

   Replace `us-east-1` with your Region.

```
$ python
>>> import boto3
>>> glue = boto3.client('glue', 'us-east-1')
>>> glue.list_blueprints()
```

Set up the preview AWS CLI

1. If you haven't already done so, install and/or update the AWS Command Line Interface (AWS CLI) on your computer. The easiest way to do this is with `pip`, the Python installer utility:

   ```
   pip install awscli --upgrade --user
   ```

   You can find complete installation instructions for the AWS CLI here: Installing the AWS Command Line Interface.
2. Download the AWS CLI wheel file from: s3://awsglue-custom-blueprints-preview-artifacts/awscli-preview-build/awscli-1.19.31-py2.py3-none-any.whl
3. Install the AWS CLI wheel file.

   ```
   python3 -m pip install awscli-1.19.31-py2.py3-none-any.whl
   ```
4. Run the `aws configure` command. Configure your AWS credentials (including access key, and secret key) and AWS Region. You can find information on configuring the AWS CLI here: Configuring the AWS CLI.
5. Test the AWS CLI. The following command should return an empty list.

   Replace `us-east-1` with your Region.

   ```
   aws glue list-blueprints --region us-east-1
   ```

Writing the blueprint code

Each blueprint project that you create must contain at a minimum the following files:

- A Python layout script that defines the workflow. The script contains a function that defines the entities (jobs and crawlers) in a workflow, and the dependencies between them.
- A configuration file, `blueprint.cfg`, which defines:

- The full path of the workflow layout definition function.
- The parameters that the blueprint accepts.

Topics

- Creating the blueprint layout script
- Creating the configuration file
- Specifying blueprint parameters

Creating the blueprint layout script

The blueprint layout script must include a function that generates the entities in your workflow. You can name this function whatever you like. AWS Glue uses the configuration file to determine the fully qualified name of the function.

Your layout function does the following:

- (Optional) Instantiates the `Job` class to create `Job` objects, and passes arguments such as `Command` and `Role`. These are job properties that you would specify if you were creating the job using the AWS Glue console or API.
- (Optional) Instantiates the `Crawler` class to create `Crawler` objects, and passes name, role, and target arguments.
- To indicate dependencies between the objects (workflow entities), passes the `DependsOn` and `WaitForDependencies` additional arguments to `Job()` and `Crawler()`. These arguments are explained later in this section.
- Instantiates the `Workflow` class to create the workflow object that is returned to AWS Glue, passing a `Name` argument, an `Entities` argument, and an optional `OnSchedule` argument. The `Entities` argument specifies all of the jobs and crawlers to include in the workflow. To see how to construct an `Entities` object, see the sample project later in this section.
- Returns the `Workflow` object.

For definitions of the `Job`, `Crawler`, and `Workflow` classes, see AWS Glue blueprint classes reference.

The layout function must accept the following input arguments.

Argument	Description
`user_params`	Python dictionary of blueprint parameter names and values. For more information, see Specifying blueprint parameters.
`system_params`	Python dictionary containing two properties: `region` and `accountId`.

Here is a sample layout generator script in a file named `Layout.py`:

```
import argparse
import sys
import os
import json
from awsglue.blueprint.workflow import *
from awsglue.blueprint.job import *
from awsglue.blueprint.crawler import *


def generate_layout(user_params, system_params):

    etl_job = Job(Name="{}_etl_job".format(user_params['WorkflowName']),
                  Command={
                      "Name": "glueetl",
                      "ScriptLocation": user_params['ScriptLocation'],
                      "PythonVersion": "2"
                  },
                  Role=user_params['PassRole'])
    post_process_job = Job(Name="{}_post_process".format(user_params['WorkflowName']),
                            Command={
                                "Name": "pythonshell",
                                "ScriptLocation": user_params['ScriptLocation'],
                                "PythonVersion": "2"
                            },
                            Role=user_params['PassRole'],
                            DependsOn={
                                etl_job: "SUCCEEDED"
                            },
                            WaitForDependencies="AND")
    sample_workflow = Workflow(Name=user_params['WorkflowName'],
                            Entities=Entities(Jobs=[etl_job, post_process_job]))
    return sample_workflow
```

The sample script imports the required blueprint libraries and includes a `generate_layout` function that generates a workflow with two jobs. This is a very simple script. A more complex script could employ additional logic and parameters to generate a workflow with many jobs and crawlers, or even a variable number of jobs and crawlers.

Using the DependsOn argument

The `DependsOn` argument is a dictionary representation of a dependency that this entity has on other entities within the workflow. It has the following form.

```
DependsOn = {dependency1 : state, dependency2 : state, ...}
```

The keys in this dictionary represent the object reference, not the name, of the entity, while the values are strings that correspond to the state to watch for. AWS Glue infers the proper triggers. For the valid states, see Condition Structure.

For example, a job might depend on the successful completion of a crawler. If you define a crawler object named `crawler2` as follows:

```
crawler2 = Crawler(Name="my_crawler", ...)
```

Then an object depending on `crawler2` would include a constructor argument such as:

```
DependsOn = {crawler2 : "SUCCEEDED"}
```

For example:

```
job1 = Job(Name="Job1", ..., DependsOn = {crawler2 : "SUCCEEDED", ...})
```

If `DependsOn` is omitted for an entity, that entity depends on the workflow start trigger.

Using the WaitForDependencies argument

The `WaitForDependencies` argument defines whether a job or crawler entity should wait until *all* entities on which it depends complete or until *any* completes.

The allowable values are "AND" or "ANY".

Using the OnSchedule argument

The `OnSchedule` argument for the `Workflow` class constructor is a `cron` expression that defines the starting trigger definition for a workflow.

If this argument is specified, AWS Glue creates a schedule trigger with the corresponding schedule. If it isn't specified, the starting trigger for the workflow is an on-demand trigger.

Creating the configuration file

The blueprint configuration file is a required file that defines the script entry point for generating the workflow, and the parameters that the blueprint accepts. The file must be named `blueprint.cfg`.

Here is a sample configuration file.

```
{
    "layoutGenerator": "DemoBlueprintProject.Layout.generate_layout",
    "parameterSpec" : {
           "WorkflowName" : {
                "type": "String",
                "collection": false
           },
           "WorkerType" : {
                "type": "String",
                "collection": false,
                "allowedValues": ["G1.X", "G2.X"],
                "defaultValue": "G1.X"
           },
           "Dpu" : {
                "type" : "Integer",
                "allowedValues" : [2, 4, 6],
                "defaultValue" : 2
           },
           "DynamoDBTableName": {
                "type": "String",
                "collection" : false
           },
           "ScriptLocation" : {
                "type": "String",
                "collection": false
```

```
      }
    }
}
```

The `layoutGenerator` property specifies the fully qualified name of the function in the script that generates the layout.

The `parameterSpec` property specifies the parameters that this blueprint accepts. For more information, see Specifying blueprint parameters.

> **Important**
>
> Your configuration file must include the workflow name as a blueprint parameter, or you must generate a unique workflow name in your layout script.

Specifying blueprint parameters

The configuration file contains blueprint parameter specifications in a `parameterSpec` JSON object. `parameterSpec` contains one or more parameter objects.

```
"parameterSpec": {
    "<parameter_name>": {
      "type": "<parameter-type>",
      "collection": true|false,
      "description": "<parameter-description>",
      "defaultValue": "<default value for the parameter if value not specified>"
      "allowedValues": "<list of allowed values>"
    },
    "<parameter_name>": {
       ...
    }
  }
```

The following are the rules for coding each parameter object:

- The parameter name and `type` are mandatory. All other properties are optional.
- If you specify the `defaultValue` property, the parameter is optional. Otherwise the parameter is mandatory and the data analyst who is creating a workflow from the blueprint must provide a value for it.

- If you set the `collection` property to `true`, the parameter can take a collection of values. Collections can be of any data type.
- If you specify `allowedValues`, the AWS Glue console displays a dropdown list of values for the data analyst to choose from when creating a workflow from the blueprint.

The following are the permitted values for `type`:

Parameter data type	Notes
`String`	-
`Integer`	-
`Double`	-
`Boolean`	Possible values are `true` and `false`. Generates a check box on the **Create a workflow from <blueprint>** page on the AWS Glue console.
`S3Uri`	Complete Amazon S3 path, beginning with `s3://`. Generates a text field and **Browse** button on the **Create a workflow from <blueprint>** page.
`S3Bucket`	Amazon S3 bucket name only. Generates a bucket picker on the **Create a workflow from <blueprint>** page.
`IAMRoleArn`	Amazon Resource Name (ARN) of an AWS Identity and Access Management (IAM) role. Generates a role picker on the **Create a workflow from <blueprint>** page.
`IAMRoleName`	Name of an IAM role. Generates a role picker on the **Create a workflow from <blueprint>** page.

Sample blueprint project

Data format conversion is a frequent extract, transform, and load (ETL) use case. In typical analytic workloads, column-based file formats like Parquet or ORC are preferred over text formats like CSV or JSON. This sample blueprint enables you to convert data from CSV/JSON/etc. into Parquet for files on Amazon S3.

This blueprint takes a list of S3 paths defined by a blueprint parameter, converts the data to Parquet format, and writes it to the S3 location specified by another blueprint parameter. The layout script creates a crawler and job for each path. The layout script also uploads the ETL script in `Conversion.py` to an S3 bucket specified by another blueprint parameter. The layout script then specifies the uploaded script as the ETL script for each job. The ZIP archive for the project contains the layout script, the ETL script, and the blueprint configuration file.

For information about more sample blueprint projects, see Blueprint samples.

The following is the layout script, in the file `Layout.py`.

```
from awsglue.blueprint.workflow import *
from awsglue.blueprint.job import *
from awsglue.blueprint.crawler import *
import boto3

s3_client = boto3.client('s3')

# Ingesting all the S3 paths as Glue table in parquet format
def generate_layout(user_params, system_params):
    #Always give the full path for the file
    with open("ConversionBlueprint/Conversion.py", "rb") as f:
        s3_client.upload_fileobj(f, user_params['ScriptsBucket'], "Conversion.py")
    etlScriptLocation = "s3://{}/Conversion.py".format(user_params['ScriptsBucket'])

    crawlers = []
    jobs = []
    workflowName = user_params['WorkflowName']
    for path in user_params['S3Paths']:
      tablePrefix = "source_"
      crawler = Crawler(Name="{}_crawler".format(workflowName),
                        Role=user_params['PassRole'],
                        DatabaseName=user_params['TargetDatabase'],
                        TablePrefix=tablePrefix,
                        Targets= {"S3Targets": [{"Path": path}]})
      crawlers.append(crawler)
      transform_job = Job(Name="{}_transform_job".format(workflowName),
                         Command={"Name": "glueetl",
                                  "ScriptLocation": etlScriptLocation,
                                  "PythonVersion": "3"},
                         Role=user_params['PassRole'],
                         DefaultArguments={"--database_name":
 user_params['TargetDatabase'],
                                            "--table_prefix": tablePrefix,
                                            "--region_name": system_params['region'],
                                            "--output_path":
 user_params['TargetS3Location']},
                         DependsOn={crawler: "SUCCEEDED"},
                         WaitForDependencies="AND")
      jobs.append(transform_job)
    conversion_workflow = Workflow(Name=workflowName, Entities=Entities(Jobs=jobs,
 Crawlers=crawlers))
    return conversion_workflow
```

The following is the corresponding blueprint configuration file `blueprint.cfg`.

```
{
    "layoutGenerator": "ConversionBlueprint.Layout.generate_layout",
    "parameterSpec" : {
        "WorkflowName" : {
            "type": "String",
            "collection": false,
            "description": "Name for the workflow."
        },
        "S3Paths" : {
            "type": "S3Uri",
            "collection": true,
            "description": "List of Amazon S3 paths for data ingestion."
        },
        "PassRole" : {
            "type": "IAMRoleName",
            "collection": false,
            "description": "Choose an IAM role to be used in running the job/crawler"
        },
        "TargetDatabase": {
            "type": "String",
            "collection" : false,
            "description": "Choose a database in the Data Catalog."
        },
        "TargetS3Location": {
            "type": "S3Uri",
            "collection" : false,
            "description": "Choose an Amazon S3 output path: ex:s3://<target_path>/."
        },
        "ScriptsBucket": {
            "type": "S3Bucket",
            "collection": false,
            "description": "Provide an S3 bucket name(in the same AWS Region) to store
 the scripts."
        }
    }
}
```

The following script in the file `Conversion.py` is the uploaded ETL script. Note that it preserves the partitioning scheme during conversion.

```
import sys
from pyspark.sql.functions import *
from pyspark.context import SparkContext
from awsglue.transforms import *
from awsglue.context import GlueContext
from awsglue.job import Job
from awsglue.utils import getResolvedOptions
import boto3

args = getResolvedOptions(sys.argv, [
    'JOB_NAME',
    'region_name',
    'database_name',
    'table_prefix',
    'output_path'])
databaseName = args['database_name']
tablePrefix = args['table_prefix']
outputPath = args['output_path']

glue = boto3.client('glue', region_name=args['region_name'])

glue_context = GlueContext(SparkContext.getOrCreate())
spark = glue_context.spark_session
job = Job(glue_context)
job.init(args['JOB_NAME'], args)

def get_tables(database_name, table_prefix):
    tables = []
    paginator = glue.get_paginator('get_tables')
    for page in paginator.paginate(DatabaseName=database_name, Expression=table_prefix
+"*"):
        tables.extend(page['TableList'])
    return tables
```

```
for table in get_tables(databaseName, tablePrefix):
    tableName = table['Name']
    partitionList = table['PartitionKeys']
    partitionKeys = []
    for partition in partitionList:
        partitionKeys.append(partition['Name'])

    # Create DynamicFrame from Catalog
    dyf = glue_context.create_dynamic_frame.from_catalog(
        name_space=databaseName,
        table_name=tableName,
        additional_options={
            'useS3ListImplementation': True
        },
        transformation_ctx='dyf'
    )

    # Resolve choice type with make_struct
    dyf = ResolveChoice.apply(
        frame=dyf,
        choice='make_struct',
        transformation_ctx='resolvechoice_' + tableName
    )

    # Drop null fields
    dyf = DropNullFields.apply(
        frame=dyf,
        transformation_ctx="dropnullfields_" + tableName
    )

    # Write DynamicFrame to S3 in glueparquet
    sink = glue_context.getSink(
        connection_type="s3",
        path=outputPath,
        enableUpdateCatalog=True,
        partitionKeys=partitionKeys
    )
    sink.setFormat("glueparquet")

    sink.setCatalogInfo(
        catalogDatabase=databaseName,
        catalogTableName=tableName[len(tablePrefix):]
    )
    sink.writeFrame(dyf)

job.commit()
```

Note

Only two Amazon S3 paths can be supplied as an input to the sample blueprint. This is because AWS Glue triggers are limited to invoking only two crawler actions.

Testing a blueprint

While you develop your code, you should perform local testing to verify that the workflow layout is correct.

Local testing doesn't generate AWS Glue jobs, crawlers, or triggers. Instead, you run the layout script locally and use the `to_json()` and `validate()` methods to print objects and find errors. These methods are available in all three classes defined in the libraries.

There are two ways to handle the `user_params` and `system_params` arguments that AWS Glue passes to your layout function. Your test-bench code can create a dictionary of sample blueprint parameter values and pass that to the layout function as the `user_params` argument. Or, you can remove the references to `user_params` and replace them with hardcoded strings.

If your code makes use of the `region` and `accountId` properties in the `system_params` argument, you can pass in your own dictionary for `system_params`.

To test a blueprint

1. Start a Python interpreter in a directory with the libraries, or load the blueprint files and the supplied libraries into your preferred integrated development environment (IDE).
2. Ensure that your code imports the supplied libraries.
3. Add code to your layout function to call `validate()` or `to_json()` on any entity or on the `Workflow` object. For example, if your code creates a `Crawler` object named `mycrawler`, you can call `validate()` as follows.

   ```
   mycrawler.validate()
   ```

   You can print `mycrawler` as follows:

```
print(mycrawler.to_json())
```

If you call `to_json` on an object, there is no need to also call `validate()`, because `to_json()` calls `validate()`.

It is most useful to call these methods on the workflow object. Assuming that your script names the workflow object `my_workflow`, validate and print the workflow object as follows.

```
print(my_workflow.to_json())
```

For more information about `to_json()` and `validate()`, see Class methods.

You can also import `pprint` and pretty-print the workflow object, as shown in the example later in this section.

4. Run the code, fix errors, and finally remove any calls to `validate()` or `to_json()`.

Example

The following example shows how to construct a dictionary of sample blueprint parameters and pass it in as the `user_params` argument to layout function `generate_compaction_workflow`. It also shows how to pretty-print the generated workflow object.

```
from pprint import pprint
from awsglue.blueprint.workflow import *
from awsglue.blueprint.job import *
from awsglue.blueprint.crawler import *

USER_PARAMS = {"WorkflowName": "compaction_workflow",
               "ScriptLocation": "s3://amzn-s3-demo-bucket/scripts/threaded-
compaction.py",
               "PassRole": "arn:aws:iam::111122223333:role/GlueRole-ETL",
               "DatabaseName": "cloudtrial",
               "TableName": "ct_cloudtrail",
               "CoalesceFactor": 4,
               "MaxThreadWorkers": 200}


def generate_compaction_workflow(user_params: dict, system_params: dict) -> Workflow:
    compaction_job = Job(Name=f"{user_params['WorkflowName']}_etl_job",
                         Command={"Name": "glueetl",
                                  "ScriptLocation": user_params['ScriptLocation'],
                                  "PythonVersion": "3"},
                         Role="arn:aws:iam::111122223333:role/
AWSGlueServiceRoleDefault",
                         DefaultArguments={"DatabaseName": user_params['DatabaseName'],
                                           "TableName": user_params['TableName'],
                                           "CoalesceFactor":
 user_params['CoalesceFactor'],
                                           "max_thread_workers":
 user_params['MaxThreadWorkers']})

    catalog_target = {"CatalogTargets": [{"DatabaseName": user_params['DatabaseName'],
 "Tables": [user_params['TableName']]}]}

    compacted_files_crawler = Crawler(Name=f"{user_params['WorkflowName']}_post_crawl",
                                      Targets = catalog_target,
                                      Role=user_params['PassRole'],
                                      DependsOn={compaction_job: "SUCCEEDED"},
                                      WaitForDependencies="AND",
                                      SchemaChangePolicy={"DeleteBehavior": "LOG"})

    compaction_workflow = Workflow(Name=user_params['WorkflowName'],
                                   Entities=Entities(Jobs=[compaction_job],

 Crawlers=[compacted_files_crawler]))
    return compaction_workflow

generated = generate_compaction_workflow(user_params=USER_PARAMS, system_params={})
gen_dict = generated.to_json()

pprint(gen_dict)
```

Publishing a blueprint

After you develop a blueprint, you must upload it to Amazon S3. You must have write permissions on the Amazon S3 bucket that you use to publish the blueprint. You must also make sure that the AWS Glue administrator, who will register the blueprint, has read access to the Amazon S3 bucket. For the suggested AWS Identity and Access Management (IAM) permissions policies for personas and roles for AWS Glue blueprints, see Permissions for personas and roles for AWS Glue blueprints.

To publish a blueprint

1. Create the necessary scripts, resources, and blueprint configuration file.
2. Add all files to a ZIP archive and upload the ZIP file to Amazon S3. Use an S3 bucket that is in the same Region as the Region in which users will register and run the blueprint.

   You can create a ZIP file from the command line using the following command.

   ```
   zip -r folder.zip folder
   ```
3. Add a bucket policy that grants read permissions to the AWS desired account. The following is a sample policy.

   JSON

   ```
   {
     "Version": "2012-10-17",
     "Statement": [
       {
         "Effect": "Allow",
         "Principal": {
           "AWS": "arn:aws:iam::111122223333:root"
         },
         "Action": "s3:GetObject",
         "Resource": "arn:aws:s3:::my-blueprints/*"
       }
     ]
   }
   ```
4. Grant the IAM `s3:GetObject` permission on the Amazon S3 bucket to the AWS Glue administrator or to whoever will be registering blueprints. For a sample policy to grant to administrators, see AWS Glue administrator permissions for blueprints.

After you have completed local testing of your blueprint, you may also want to test a blueprint on AWS Glue. To test a blueprint on AWS Glue, it must be registered. You can limit who sees the registered blueprint using IAM authorization, or by using separate testing accounts.

See also:

- Registering a blueprint in AWS Glue

AWS Glue blueprint classes reference

The libraries for AWS Glue blueprints define three classes that you use in your workflow layout script: `Job`, `Crawler`, and `Workflow`.

Topics

- Job class
- Crawler class
- Workflow class
- Class methods

Job class

The `Job` class represents an AWS Glue ETL job.

Mandatory constructor arguments

The following are mandatory constructor arguments for the `Job` class.

Argument name	Type	Description
`Name`	`str`	Name to assign to the job. AWS Glue adds a randomly generated suffix to the name to distinguish the job from those created by other blueprint runs.
`Role`	`str`	Amazon Resource Name (ARN) of the role that the job should assume while executing.
`Command`	`dict`	Job command, as specified in the JobCommand structure in the API documentation.

Optional constructor arguments

The following are optional constructor arguments for the `Job` class.

Argument name	Type	Description
`DependsOn`	`dict`	List of workflow entities that the job depends on. For more information, see Using the DependsOn argument.
`WaitForDependencies`	`str`	Indicates whether the job should wait until *all* entities on which it depends complete before executing or until *any* completes. For more information, see Using the WaitForDependencies argument. Omit if the job depends on only one entity.
(Job properties)	-	Any of the job properties listed in Job structure in the AWS Glue API documentation (except `CreatedOn` and `LastModifiedOn`).

Crawler class

The `Crawler` class represents an AWS Glue crawler.

Mandatory constructor arguments

The following are mandatory constructor arguments for the `Crawler` class.

Argument name	Type	Description
`Name`	`str`	Name to assign to the crawler. AWS Glue adds a randomly generated suffix to the name to distinguish the crawler from those created by other blueprint runs.
`Role`	`str`	ARN of the role that the crawler should assume while running.

Argument name	Type	Description
`Targets`	`dict`	Collection of targets to crawl. `Targets` class constructor arguments are defined in the CrawlerTargets structure in the API documentation. All `Targets` constructor arguments are optional, but you must pass at least one.

Optional constructor arguments

The following are optional constructor arguments for the `Crawler` class.

Argument name	Type	Description
`DependsOn`	`dict`	List of workflow entities that the crawler depends on. For more information, see Using the DependsOn argument.
`WaitForDependencies`	`str`	Indicates whether the crawler should wait until *all* entities on which it depends complete before running or until *any* completes. For more information, see Using the WaitForDependencies argument. Omit if the crawler depends on only one entity.
(Crawler properties)	-	Any of the crawler properties listed in Crawler structure in the AWS Glue API documentation, with the following exceptions: • `State` • `CrawlElapsedTime` • `CreationTime` • `LastUpdated` • `LastCrawl` • `Version`

Workflow class

The `Workflow` class represents an AWS Glue workflow. The workflow layout script returns a `Workflow` object. AWS Glue creates a workflow based on this object.

Mandatory constructor arguments

The following are mandatory constructor arguments for the `Workflow` class.

Argument name	Type	Description
`Name`	`str`	Name to assign to the workflow.
`Entities`	`Entities`	A collection of entities (jobs and crawlers) to include in the workflow. The `Entities` class constructor accepts a `Jobs` argument, which is a list of `Job` objects, and a `Crawlers` argument, which is a list of `Crawler` objects.

Optional constructor arguments

The following are optional constructor arguments for the `Workflow` class.

Argument name	Type	Description
`Description`	`str`	See Workflow structure.
`DefaultRunProperties`	`dict`	See Workflow structure.
`OnSchedule`	`str`	A `cron` expression.

Class methods

All three classes include the following methods.

validate()

Validates the properties of the object and if errors are found, outputs a message and exits. Generates no output if there are no errors. For the `Workflow` class, calls itself on every entity in the workflow.

to_json()

Serializes the object to JSON. Also calls `validate()`. For the `Workflow` class, the JSON object includes job and crawler lists, and a list of triggers generated by the job and crawler dependency specifications.

Blueprint samples

There are a number of sample blueprint projects available on the AWS Glue blueprint Github repository. These samples are for reference only and are not intended for production use.

The titles of the sample projects are:

- Compaction: this blueprint creates a job that compacts input files into larger chunks based on desired file size.
- Conversion: this blueprint converts input files in various standard file formats into Apache Parquet format, which is optimized for analytic workloads.
- Crawling Amazon S3 locations: this blueprint crawls multiple Amazon S3 locations to add metadata tables to the Data Catalog.
- Custom connection to Data Catalog: this blueprint accesses data stores using AWS Glue custom connectors, reads the records, and populates the table definitions in the AWS Glue Data Catalog based on the record schema.
- Encoding: this blueprint converts your non-UTF files into UTF encoded files.
- Partitioning: this blueprint creates a partitioning job that places output files into partitions based on specific partition keys.
- Importing Amazon S3 data into a DynamoDB table: this blueprint imports data from Amazon S3 into a DynamoDB table.
- Standard table to governed: this blueprint imports an AWS Glue Data Catalog table into a Lake Formation table.

Registering a blueprint in AWS Glue

After the AWS Glue developer has coded the blueprint and uploaded a ZIP archive to Amazon Simple Storage Service (Amazon S3), an AWS Glue administrator must register the blueprint. Registering the blueprint makes it available for use.

When you register a blueprint, AWS Glue copies the blueprint archive to a reserved Amazon S3 location. You can then delete the archive from the upload location.

To register a blueprint, you need read permissions on the Amazon S3 location that contains the uploaded archive. You also need the AWS Identity and Access Management (IAM) permission `glue:CreateBlueprint`. For the suggested permissions for an AWS Glue administrator who must register, view, and maintain blueprints, see AWS Glue administrator permissions for blueprints.

You can register a blueprint by using the AWS Glue console, AWS Glue API, or AWS Command Line Interface (AWS CLI).

To register a blueprint (console)

1. Ensure that you have read permissions (`s3:GetObject`) on the blueprint ZIP archive in Amazon S3.
2. Open the AWS Glue console at https://console.aws.amazon.com/glue/.

   Sign in as a user that has permissions to register a blueprint. Switch to the same AWS Region as the Amazon S3 bucket that contains the blueprint ZIP archive.
3. In the navigation pane, choose **blueprints**. Then on the **blueprints** page, choose **Add blueprint**.
4. Enter a blueprint name and optional description.
5. For **ZIP archive location (S3)**, enter the Amazon S3 path of the uploaded blueprint ZIP archive. Include the archive file name in the path and begin the path with `s3://`.
6. (Optional) Add tag one or more tags.
7. Choose **Add blueprint**.

   The **blueprints** page returns and shows that the blueprint status is `CREATING`. Choose the refresh button until the status changes to `ACTIVE` or `FAILED`.
8. If the status is `FAILED`, select the blueprint, and on the **Actions** menu, choose **View**.

The detail page shows the reason for the failure. If the error message is "Unable to access object at location..." or "Access denied on object at location...", review the following requirements:

- The user that you are signed in as must have read permission on the blueprint ZIP archive in Amazon S3.
- The Amazon S3 bucket that contains the ZIP archive must have a bucket policy that grants read permission on the object to your AWS account ID. For more information, see Developing blueprints in AWS Glue.
- The Amazon S3 bucket that you're using must be in the same Region as the Region that you're signed into on the console.

9. Ensure that data analysts have permissions on the blueprint.

   The suggested IAM policy for data analysts is shown in Data analyst permissions for blueprints. This policy grants `glue:GetBlueprint` on any resource. If your policy is more fine-grained at the resource level, then grant data analysts permissions on this newly created resource.

To register a blueprint (AWS CLI)

1. Enter the following command.

```
aws glue create-blueprint --name <blueprint-name> [--description <description>] --
blueprint-location s3://<s3-path>/<archive-filename>
```

2. Enter the following command to check the blueprint status. Repeat the command until the status goes to `ACTIVE` or `FAILED`.

```
aws glue get-blueprint --name <blueprint-name>
```

   If the status is `FAILED` and the error message is "Unable to access object at location..." or "Access denied on object at location...", review the following requirements:

   - The user that you are signed in as must have read permission on the blueprint ZIP archive in Amazon S3.
   - The Amazon S3 bucket containing the ZIP archive must have a bucket policy that grants read permission on the object to your AWS account ID. For more information, see Publishing a blueprint.

- The Amazon S3 bucket that you're using must be in the same Region as the Region that you're signed into on the console.

See also:

- Overview of blueprints in AWS Glue

Viewing blueprints in AWS Glue

View a blueprint to review the blueprint description, status, and parameter specifications, and to download the blueprint ZIP archive.

You can view a blueprint by using the AWS Glue console, AWS Glue API, or AWS Command Line Interface (AWS CLI).

To view a blueprint (console)

1. Open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, choose **blueprints**.
3. On the **blueprints** page, select a blueprint. Then on the **Actions** menu, choose **View**.

To view a blueprint (AWS CLI)

- Enter the following command to view just the blueprint name, description, and status. Replace *<blueprint-name>* with the name of the blueprint to view.

  ```
  aws glue get-blueprint --name <blueprint-name>
  ```

  The command output looks something like the following.

  ```
  {
      "Blueprint": {
          "Name": "myDemoBP",
          "CreatedOn": 1587414516.92,
          "LastModifiedOn": 1587428838.671,
          "BlueprintLocation": "s3://amzn-s3-demo-bucket1/demo/
  DemoBlueprintProject.zip",
  ```

```
        "Status": "ACTIVE"
    }
}
```

Enter the following command to also view the parameter specifications.

```
aws glue get-blueprint --name <blueprint-name>  --include-parameter-spec
```

The command output looks something like the following.

```
{
    "Blueprint": {
        "Name": "myDemoBP",
        "CreatedOn": 1587414516.92,
        "LastModifiedOn": 1587428838.671,
        "ParameterSpec": "{\"WorkflowName\":{\"type\":\"String\",\"collection
\":false,\"description\":null,\"defaultValue\":null,\"allowedValues\":null},
\"PassRole\":{\"type\":\"String\",\"collection\":false,\"description\":null,
\"defaultValue\":null,\"allowedValues\":null},\"DynamoDBTableName\":{\"type
\":\"String\",\"collection\":false,\"description\":null,\"defaultValue\":null,
\"allowedValues\":null},\"ScriptLocation\":{\"type\":\"String\",\"collection
\":false,\"description\":null,\"defaultValue\":null,\"allowedValues\":null}}",
        "BlueprintLocation": "s3://awsexamplebucket1/demo/
DemoBlueprintProject.zip",
        "Status": "ACTIVE"
    }
}
```

Add the `--include-blueprint` argument to include a URL in the output that you can paste into your browser to download the blueprint ZIP archive that AWS Glue stored.

See also:

- Overview of blueprints in AWS Glue

Updating a blueprint in AWS Glue

You can update a blueprint if you have a revised layout script, a revised set of blueprint parameters, or revised supporting files. Updating a blueprint creates a new version.

Updating a blueprint doesn't affect existing workflows created from the blueprint.

You can update a blueprint by using the AWS Glue console, AWS Glue API, or AWS Command Line Interface (AWS CLI).

The following procedure assumes that the AWS Glue developer has created and uploaded an updated blueprint ZIP archive to Amazon S3.

To update a blueprint (console)

1. Ensure that you have read permissions (`s3:GetObject`) on the blueprint ZIP archive in Amazon S3.
2. Open the AWS Glue console at https://console.aws.amazon.com/glue/.

   Sign in as a user that has permissions to update a blueprint. Switch to the same AWS Region as the Amazon S3 bucket that contains the blueprint ZIP archive.
3. In the navigation pane, choose **blueprints**.
4. On the **blueprints** page, select a blueprint, and on the **Actions** menu, choose **Edit**.
5. On the **Edit a blueprint** page, update the blueprint **Description** or **ZIP archive location (S3)**. Be sure to include the archive file name in the path.
6. Choose **Save**.

   The **blueprints** page returns and shows that the blueprint status is `UPDATING`. Choose the refresh button until the status changes to `ACTIVE` or `FAILED`.
7. If the status is `FAILED`, select the blueprint, and on the **Actions** menu, choose **View**.

   The detail page shows the reason for the failure. If the error message is "Unable to access object at location..." or "Access denied on object at location...", review the following requirements:

   - The user that you are signed in as must have read permission on the blueprint ZIP archive in Amazon S3.

- The Amazon S3 bucket that contains the ZIP archive must have a bucket policy that grants read permission on the object to your AWS account ID. For more information, see Publishing a blueprint.
- The Amazon S3 bucket that you're using must be in the same Region as the Region that you're signed into on the console.

> **Note**
>
> If the update fails, the next blueprint run uses the latest version of the blueprint that was successfully registered or updated.

To update a blueprint (AWS CLI)

1. Enter the following command.

   ```
   aws glue update-blueprint --name <blueprint-name> [--description <description>] --
   blueprint-location s3://<s3-path>/<archive-filename>
   ```

2. Enter the following command to check the blueprint status. Repeat the command until the status goes to `ACTIVE` or `FAILED`.

   ```
   aws glue get-blueprint --name <blueprint-name>
   ```

   If the status is `FAILED` and the error message is "Unable to access object at location..." or "Access denied on object at location...", review the following requirements:

   - The user that you are signed in as must have read permission on the blueprint ZIP archive in Amazon S3.
   - The Amazon S3 bucket containing the ZIP archive must have a bucket policy that grants read permission on the object to your AWS account ID. For more information, see Publishing a blueprint.
   - The Amazon S3 bucket that you're using must be in the same Region as the Region that you're signed into on the console.

See also

- Overview of blueprints in AWS Glue

Creating a workflow from a blueprint in AWS Glue

You can create an AWS Glue workflow manually, adding one component at a time, or you can create a workflow from an AWS Glue blueprint. AWS Glue includes blueprints for common use cases. Your AWS Glue developers can create additional blueprints.

Important

Limit the total number of jobs, crawlers, and triggers within a workflow to 100 or less. If you include more than 100, you might get errors when trying to resume or stop workflow runs.

When you use a blueprint, you can quickly generate a workflow for a specific use case based on the generalized use case defined by the blueprint. You define the specific use case by providing values for the blueprint parameters. For example, a blueprint that partitions a dataset could have the Amazon S3 source and target paths as parameters.

AWS Glue creates a workflow from a blueprint by *running* the blueprint. The blueprint run saves the parameter values that you supplied, and is used to track the progress and outcome of the creation of the workflow and its components. When troubleshooting a workflow, you can view the blueprint run to determine the blueprint parameter values that were used to create a workflow.

To create and view workflows, you require certain IAM permissions. For a suggested IAM policy, see Data analyst permissions for blueprints.

You can create a workflow from a blueprint by using the AWS Glue console, AWS Glue API, or AWS Command Line Interface (AWS CLI).

To create a workflow from a blueprint (console)

1. Open the AWS Glue console at https://console.aws.amazon.com/glue/.

   Sign in as a user that has permissions to create a workflow.

2. In the navigation pane, choose **blueprints**.
3. Select a blueprint, and on the **Actions** menu, choose **Create workflow**.
4. On the **Create a workflow from <blueprint-name>** page, enter the following information:

   Blueprint parameters

      These vary depending on the blueprint design. For questions about the parameters, see the developer. blueprints typically include a parameter for the workflow name.

   IAM role

      The role that AWS Glue assumes to create the workflow and its components. The role must have permissions to create and delete workflows, jobs, crawlers, and triggers. For a suggested policy for the role, see Permissions for blueprint roles.
5. Choose **Submit**.

   The **Blueprint Details** page appears, showing a list of blueprint runs at the bottom.
6. In the blueprint runs list, check the topmost blueprint run for workflow creation status.

   The initial status is `RUNNING`. Choose the refresh button until the status goes to `SUCCEEDED` or `FAILED`.
7. Do one of the following:

   - If the completion status is `SUCCEEDED`, you can go to the **Workflows** page, select the newly created workflow, and run it. Before running the workflow, you can review the design graph.
   - If the completion status is `FAILED`, select the blueprint run, and on the **Actions** menu, choose **View** to see the error message.

For more information on workflows and blueprints, see the following topics.

- Overview of workflows in AWS Glue
- Updating a blueprint in AWS Glue
- Creating and building out a workflow manually in AWS Glue

Viewing blueprint runs in AWS Glue

View a blueprint run to see the following information:

- Name of the workflow that was created.
- blueprint parameter values that were used to create the workflow.
- Status of the workflow creation operation.

You can view a blueprint run by using the AWS Glue console, AWS Glue API, or AWS Command Line Interface (AWS CLI).

To view a blueprint run (console)

1. Open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane, choose **blueprints**.
3. On the **blueprints** page, select a blueprint. Then on the **Actions** menu, choose **View**.
4. At the bottom of the **Blueprint Details** page, select a blueprint run, and on the **Actions** menu, choose **View**.

To view a blueprint run (AWS CLI)

- Enter the following command. Replace *<blueprint-name>* with the name of the blueprint. Replace *<blueprint-run-id>* with the blueprint run ID.

```
aws glue get-blueprint-run --blueprint-name <blueprint-name> --run-id <blueprint-run-id>
```

See also:

- [Overview of blueprints in AWS Glue]()

AWS CloudFormation for AWS Glue

AWS CloudFormation is a service that can create many AWS resources. AWS Glue provides API operations to create objects in the AWS Glue Data Catalog. However, it might be more convenient to define and create AWS Glue objects and other related AWS resource objects in an AWS CloudFormation template file. Then you can automate the process of creating the objects.

AWS CloudFormation provides a simplified syntax—either JSON (JavaScript Object Notation) or YAML (YAML Ain't Markup Language)—to express the creation of AWS resources. You can use AWS CloudFormation templates to define Data Catalog objects such as databases, tables, partitions, crawlers, classifiers, and connections. You can also define ETL objects such as jobs, triggers, and development endpoints. You create a template that describes all the AWS resources you want, and AWS CloudFormation takes care of provisioning and configuring those resources for you.

For more information, see What Is AWS CloudFormation? and Working with AWS CloudFormation Templates in the *AWS CloudFormation User Guide*.

If you plan to use AWS CloudFormation templates that are compatible with AWS Glue, as an administrator, you must grant access to AWS CloudFormation and to the AWS services and actions on which it depends. To grant permissions to create AWS CloudFormation resources, attach the following policy to users that work with AWS CloudFormation:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "cloudformation:*"
      ],
      "Resource": "*"
    }
  ]
}
```

The following table contains the actions that an AWS CloudFormation template can perform on your behalf. It includes links to information about the AWS resource types and their property types that you can add to an AWS CloudFormation template.

AWS Glue resource	AWS CloudFormation template	AWS Glue samples
Classifier	AWS::Glue::Classifier	Grok classifier, JSON classifier, XML classifier
Connection	AWS::Glue::Connection	MySQL connection
Crawler	AWS::Glue::Crawler	Amazon S3 crawler, MySQL crawler
Database	AWS::Glue::Database	Empty database, Database with tables
Development endpoint	AWS::Glue::DevEndpoint	Development endpoint
Job	AWS::Glue::Job	Amazon S3 job, JDBC job
Machine learning transform	AWS::Glue::MLTransform	Machine learning transform
Data quality ruleset	AWS::Glue::DataQualityRuleset	Data quality ruleset, Data quality ruleset with EventBridge scheduler
Partition	AWS::Glue::Partition	Partitions of a table
Table	AWS::Glue::Table	Table in a database
Trigger	AWS::Glue::Trigger	On-demand trigger, Scheduled trigger, Conditional trigger

To get started, use the following sample templates and customize them with your own metadata. Then use the AWS CloudFormation console to create an AWS CloudFormation stack to add objects to AWS Glue and any associated services. Many fields in an AWS Glue object are optional. These templates illustrate the fields that are required or are necessary for a working and functional AWS Glue object.

An AWS CloudFormation template can be in either JSON or YAML format. In these examples, YAML is used for easier readability. The examples contain comments (#) to describe the values that are defined in the templates.

AWS CloudFormation templates can include a `Parameters` section. This section can be changed in the sample text or when the YAML file is submitted to the AWS CloudFormation console to create a stack. The `Resources` section of the template contains the definition of AWS Glue and related objects. AWS CloudFormation template syntax definitions might contain properties that include more detailed property syntax. Not all properties might be required to create an AWS Glue object. These samples show example values for common properties to create an AWS Glue object.

Sample AWS CloudFormation template for an AWS Glue database

An AWS Glue database in the Data Catalog contains metadata tables. The database consists of very few properties and can be created in the Data Catalog with an AWS CloudFormation template. The following sample template is provided to get you started and to illustrate the use of AWS CloudFormation stacks with AWS Glue. The only resource created by the sample template is a database named `cfn-mysampledatabase`. You can change it by editing the text of the sample or changing the value on the AWS CloudFormation console when you submit the YAML.

The following shows example values for common properties to create an AWS Glue database. For more information about the AWS CloudFormation database template for AWS Glue, see AWS::Glue::Database.

```
---
AWSTemplateFormatVersion: '2010-09-09'
# Sample CloudFormation template in YAML to demonstrate creating a database named
 mysampledatabase
# The metadata created in the Data Catalog points to the flights public S3 bucket
#
# Parameters section contains names that are substituted in the Resources section
# These parameters are the names the resources created in the Data Catalog
Parameters:
  CFNDatabaseName:
    Type: String
    Default: cfn-mysampledatabse

# Resources section defines metadata for the Data Catalog
Resources:
# Create an AWS Glue database
  CFNDatabaseFlights:
    Type: AWS::Glue::Database
    Properties:
      # The database is created in the Data Catalog for your account
      CatalogId: !Ref AWS::AccountId
      DatabaseInput:
        # The name of the database is defined in the Parameters section above
        Name: !Ref CFNDatabaseName
        Description: Database to hold tables for flights data
        LocationUri: s3://crawler-public-us-east-1/flight/2016/csv/
        #Parameters: Leave AWS database parameters blank
```

Sample AWS CloudFormation template for an AWS Glue database, table, and partition

An AWS Glue table contains the metadata that defines the structure and location of data that you want to process with your ETL scripts. Within a table, you can define partitions to parallelize the processing of your data. A partition is a chunk of data that you defined with a key. For example, using month as a key, all the data for January is contained in the same partition. In AWS Glue, databases can contain tables, and tables can contain partitions.

The following sample shows how to populate a database, a table, and partitions using an AWS CloudFormation template. The base data format is `csv` and delimited by a comma (,). Because a database must exist before it can contain a table, and a table must exist before partitions can be created, the template uses the `DependsOn` statement to define the dependency of these objects when they are created.

The values in this sample define a table that contains flight data from a publicly available Amazon S3 bucket. For illustration, only a few columns of the data and one partitioning key are defined. Four partitions are also defined in the Data Catalog. Some fields to describe the storage of the base data are also shown in the `StorageDescriptor` fields.

```

---
AWSTemplateFormatVersion: '2010-09-09'
```

```
# Sample CloudFormation template in YAML to demonstrate creating a database, a table,
 and partitions
# The metadata created in the Data Catalog points to the flights public S3 bucket
#
# Parameters substituted in the Resources section
# These parameters are names of the resources created in the Data Catalog
Parameters:
  CFNDatabaseName:
    Type: String
    Default: cfn-database-flights-1
  CFNTableName1:
    Type: String
    Default: cfn-manual-table-flights-1
# Resources to create metadata in the Data Catalog
Resources:
###
# Create an AWS Glue database
  CFNDatabaseFlights:
    Type: AWS::Glue::Database
    Properties:
      CatalogId: !Ref AWS::AccountId
      DatabaseInput:
        Name: !Ref CFNDatabaseName
        Description: Database to hold tables for flights data
###
# Create an AWS Glue table
  CFNTableFlights:
    # Creating the table waits for the database to be created
    DependsOn: CFNDatabaseFlights
    Type: AWS::Glue::Table
    Properties:
      CatalogId: !Ref AWS::AccountId
      DatabaseName: !Ref CFNDatabaseName
      TableInput:
        Name: !Ref CFNTableName1
        Description: Define the first few columns of the flights table
        TableType: EXTERNAL_TABLE
        Parameters: {
    "classification": "csv"
  }
#       ViewExpandedText: String
        PartitionKeys:
        # Data is partitioned by month
        - Name: mon
```

```
          Type: bigint
        StorageDescriptor:
          OutputFormat: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
          Columns:
          - Name: year
            Type: bigint
          - Name: quarter
            Type: bigint
          - Name: month
            Type: bigint
          - Name: day_of_month
            Type: bigint
          InputFormat: org.apache.hadoop.mapred.TextInputFormat
          Location: s3://crawler-public-us-east-1/flight/2016/csv/
          SerdeInfo:
            Parameters:
              field.delim: ","
            SerializationLibrary: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
# Partition 1
# Create an AWS Glue partition
  CFNPartitionMon1:
    DependsOn: CFNTableFlights
    Type: AWS::Glue::Partition
    Properties:
      CatalogId: !Ref AWS::AccountId
      DatabaseName: !Ref CFNDatabaseName
      TableName: !Ref CFNTableName1
      PartitionInput:
        Values:
        - 1
        StorageDescriptor:
          OutputFormat: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
          Columns:
          - Name: mon
            Type: bigint
          InputFormat: org.apache.hadoop.mapred.TextInputFormat
          Location: s3://crawler-public-us-east-1/flight/2016/csv/mon=1/
          SerdeInfo:
            Parameters:
              field.delim: ","
            SerializationLibrary: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
# Partition 2
# Create an AWS Glue partition
  CFNPartitionMon2:
    DependsOn: CFNTableFlights
    Type: AWS::Glue::Partition
    Properties:
      CatalogId: !Ref AWS::AccountId
      DatabaseName: !Ref CFNDatabaseName
      TableName: !Ref CFNTableName1
      PartitionInput:
        Values:
        - 2
        StorageDescriptor:
          OutputFormat: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
          Columns:
          - Name: mon
            Type: bigint
          InputFormat: org.apache.hadoop.mapred.TextInputFormat
          Location: s3://crawler-public-us-east-1/flight/2016/csv/mon=2/
          SerdeInfo:
            Parameters:
              field.delim: ","
            SerializationLibrary: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
# Partition 3
# Create an AWS Glue partition
  CFNPartitionMon3:
    DependsOn: CFNTableFlights
    Type: AWS::Glue::Partition
    Properties:
      CatalogId: !Ref AWS::AccountId
      DatabaseName: !Ref CFNDatabaseName
      TableName: !Ref CFNTableName1
      PartitionInput:
        Values:
        - 3
        StorageDescriptor:
          OutputFormat: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
          Columns:
          - Name: mon
            Type: bigint
          InputFormat: org.apache.hadoop.mapred.TextInputFormat
          Location: s3://crawler-public-us-east-1/flight/2016/csv/mon=3/
          SerdeInfo:
            Parameters:
              field.delim: ","
            SerializationLibrary: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
# Partition 4
# Create an AWS Glue partition
  CFNPartitionMon4:
    DependsOn: CFNTableFlights
    Type: AWS::Glue::Partition
    Properties:
      CatalogId: !Ref AWS::AccountId
      DatabaseName: !Ref CFNDatabaseName
      TableName: !Ref CFNTableName1
      PartitionInput:
        Values:
        - 4
        StorageDescriptor:
          OutputFormat: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
          Columns:
          - Name: mon
            Type: bigint
          InputFormat: org.apache.hadoop.mapred.TextInputFormat
          Location: s3://crawler-public-us-east-1/flight/2016/csv/mon=4/
          SerdeInfo:
            Parameters:
              field.delim: ","
            SerializationLibrary: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
```

Sample AWS CloudFormation template for an AWS Glue grok classifier

An AWS Glue classifier determines the schema of your data. One type of custom classifier uses a grok pattern to match your data. If the pattern matches, then the custom classifier is used to create your table's schema and set the `classification` to the value set in the classifier definition.

This sample creates a classifier that creates a schema with one column named `message` and sets the classification to `greedy`.

```
---
AWSTemplateFormatVersion: '2010-09-09'
# Sample CFN YAML to demonstrate creating a classifier
#
# Parameters section contains names that are substituted in the Resources section
# These parameters are the names the resources created in the Data Catalog
```

```
Parameters:

# The name of the classifier to be created
  CFNClassifierName:
    Type: String
    Default: cfn-classifier-grok-one-column-1

#
#
# Resources section defines metadata for the Data Catalog
Resources:
# Create classifier that uses grok pattern to put all data in one column and classifies
 it as "greedy".
  CFNClassifierFlights:
    Type: AWS::Glue::Classifier
    Properties:
      GrokClassifier:
        #Grok classifier that puts all data in one column
        Name: !Ref CFNClassifierName
        Classification: greedy

        GrokPattern: "%{GREEDYDATA:message}"
        #CustomPatterns: none
```

Sample AWS CloudFormation template for an AWS Glue JSON classifier

An AWS Glue classifier determines the schema of your data. One type of custom classifier uses a `JsonPath` string defining the JSON data for the classifier to classify. AWS Glue supports a subset of the operators for `JsonPath`, as described in Writing JsonPath Custom Classifiers.

If the pattern matches, then the custom classifier is used to create your table's schema.

This sample creates a classifier that creates a schema with each record in the `Records3` array in an object.

```

---
AWSTemplateFormatVersion: '2010-09-09'
# Sample CFN YAML to demonstrate creating a JSON classifier
#
```

```
# Parameters section contains names that are substituted in the Resources section
# These parameters are the names the resources created in the Data Catalog
Parameters:

# The name of the classifier to be created
  CFNClassifierName:
    Type: String
    Default: cfn-classifier-json-one-column-1

#
#
# Resources section defines metadata for the Data Catalog
Resources:
# Create classifier that uses a JSON pattern.
  CFNClassifierFlights:
    Type: AWS::Glue::Classifier
    Properties:
      JSONClassifier:
        #JSON classifier
        Name: !Ref CFNClassifierName
        JsonPath: $.Records3[*]
```

Sample AWS CloudFormation template for an AWS Glue XML classifier

An AWS Glue classifier determines the schema of your data. One type of custom classifier specifies an XML tag to designate the element that contains each record in an XML document that is being parsed. If the pattern matches, then the custom classifier is used to create your table's schema and set the `classification` to the value set in the classifier definition.

This sample creates a classifier that creates a schema with each record in the `Record` tag and sets the classification to `XML`.

```
---
AWSTemplateFormatVersion: '2010-09-09'
# Sample CFN YAML to demonstrate creating an XML classifier
#
# Parameters section contains names that are substituted in the Resources section
# These parameters are the names the resources created in the Data Catalog
```

```
Parameters:

# The name of the classifier to be created
  CFNClassifierName:
    Type: String
    Default: cfn-classifier-xml-one-column-1

#
#
# Resources section defines metadata for the Data Catalog
Resources:
# Create classifier that uses the XML pattern and classifies it as "XML".
  CFNClassifierFlights:
    Type: AWS::Glue::Classifier
    Properties:
      XMLClassifier:
        #XML classifier
        Name: !Ref CFNClassifierName
        Classification: XML
        RowTag: <Records>
```

Sample AWS CloudFormation template for an AWS Glue crawler for Amazon S3

An AWS Glue crawler creates metadata tables in your Data Catalog that correspond to your data. You can then use these table definitions as sources and targets in your ETL jobs.

This sample creates a crawler, the required IAM role, and an AWS Glue database in the Data Catalog. When this crawler is run, it assumes the IAM role and creates a table in the database for the public flights data. The table is created with the prefix "`cfn_sample_1_`". The IAM role created by this template allows global permissions; you might want to create a custom role. No custom classifiers are defined by this classifier. AWS Glue built-in classifiers are used by default.

When you submit this sample to the AWS CloudFormation console, you must confirm that you want to create the IAM role.

```

---
AWSTemplateFormatVersion: '2010-09-09'
# Sample CFN YAML to demonstrate creating a crawler
#
```

```
# Parameters section contains names that are substituted in the Resources section
# These parameters are the names the resources created in the Data Catalog
Parameters:

# The name of the crawler to be created
  CFNCrawlerName:
    Type: String
    Default: cfn-crawler-flights-1
  CFNDatabaseName:
    Type: String
    Default: cfn-database-flights-1
  CFNTablePrefixName:
    Type: String
    Default: cfn_sample_1_
#
#
# Resources section defines metadata for the Data Catalog
Resources:
#Create IAM Role assumed by the crawler. For demonstration, this role is given all
 permissions.
  CFNRoleFlights:
    Type: AWS::IAM::Role
    Properties:
      AssumeRolePolicyDocument:
        Version: "2012-10-17"
        Statement:
          -
            Effect: "Allow"
            Principal:
              Service:
                - "glue.amazonaws.com"
            Action:
              - "sts:AssumeRole"
      Path: "/"
      Policies:
        -
          PolicyName: "root"
          PolicyDocument:
            Version: "2012-10-17"
            Statement:
              -
                Effect: "Allow"
                Action: "*"
                Resource: "*"
 # Create a database to contain tables created by the crawler
  CFNDatabaseFlights:
    Type: AWS::Glue::Database
    Properties:
      CatalogId: !Ref AWS::AccountId
      DatabaseInput:
        Name: !Ref CFNDatabaseName
        Description: "AWS Glue container to hold metadata tables for the flights 
 crawler"
 #Create a crawler to crawl the flights data on a public S3 bucket
  CFNCrawlerFlights:
    Type: AWS::Glue::Crawler
    Properties:
      Name: !Ref CFNCrawlerName
      Role: !GetAtt CFNRoleFlights.Arn
      #Classifiers: none, use the default classifier
      Description: AWS Glue crawler to crawl flights data
      #Schedule: none, use default run-on-demand
      DatabaseName: !Ref CFNDatabaseName
      Targets:
        S3Targets:
          # Public S3 bucket with the flights data
          - Path: "s3://crawler-public-us-east-1/flight/2016/csv"
      TablePrefix: !Ref CFNTablePrefixName
      SchemaChangePolicy:
        UpdateBehavior: "UPDATE_IN_DATABASE"
        DeleteBehavior: "LOG"
      Configuration: "{\"Version\":1.0,\"CrawlerOutput\":{\"Partitions\":
{\"AddOrUpdateBehavior\":\"InheritFromTable\"},\"Tables\":{\"AddOrUpdateBehavior\":
\"MergeNewColumns\"}}}"
```

Sample AWS CloudFormation template for an AWS Glue connection

An AWS Glue connection in the Data Catalog contains the JDBC and network information that is required to connect to a JDBC database. This information is used when you connect to a JDBC database to crawl or run ETL jobs.

This sample creates a connection to an Amazon RDS MySQL database named `devdb`. When this connection is used, an IAM role, database credentials, and network connection values must also be supplied. See the details of necessary fields in the template.

```
---
AWSTemplateFormatVersion: '2010-09-09'
# Sample CFN YAML to demonstrate creating a connection
#
# Parameters section contains names that are substituted in the Resources section
# These parameters are the names the resources created in the Data Catalog
Parameters:

# The name of the connection to be created
  CFNConnectionName:
    Type: String
    Default: cfn-connection-mysql-flights-1
  CFNJDBCString:
    Type: String
    Default: "jdbc:mysql://xxx-mysql.yyyyyyyyyyyyyy.us-east-1.rds.amazonaws.com:3306/
devdb"
  CFNJDBCUser:
    Type: String
    Default: "master"
  CFNJDBCPassword:
    Type: String
    Default: "12345678"
    NoEcho: true
#
#
# Resources section defines metadata for the Data Catalog
Resources:
  CFNConnectionMySQL:
    Type: AWS::Glue::Connection
    Properties:
      CatalogId: !Ref AWS::AccountId
      ConnectionInput:
        Description: "Connect to MySQL database."
        ConnectionType: "JDBC"
        #MatchCriteria: none
        PhysicalConnectionRequirements:
          AvailabilityZone: "us-east-1d"
          SecurityGroupIdList:
           - "sg-7d52b812"
          SubnetId: "subnet-84f326ee"
        ConnectionProperties: {
          "JDBC_CONNECTION_URL": !Ref CFNJDBCString,
          "USERNAME": !Ref CFNJDBCUser,
          "PASSWORD": !Ref CFNJDBCPassword
        }
        Name: !Ref CFNConnectionName
```

Sample AWS CloudFormation template for an AWS Glue crawler for JDBC

An AWS Glue crawler creates metadata tables in your Data Catalog that correspond to your data. You can then use these table definitions as sources and targets in your ETL jobs.

This sample creates a crawler, required IAM role, and an AWS Glue database in the Data Catalog. When this crawler is run, it assumes the IAM role and creates a table in the database for the public flights data that has been stored in a MySQL database. The table is created with the prefix "`cfn_jdbc_1_`". The IAM role created by this template allows global permissions; you might want to create a custom role. No custom classifiers can be defined for JDBC data. AWS Glue built-in classifiers are used by default.

When you submit this sample to the AWS CloudFormation console, you must confirm that you want to create the IAM role.

```

---
AWSTemplateFormatVersion: '2010-09-09'
# Sample CFN YAML to demonstrate creating a crawler
#
# Parameters section contains names that are substituted in the Resources section
# These parameters are the names the resources created in the Data Catalog
Parameters:

# The name of the crawler to be created
  CFNCrawlerName:
    Type: String
    Default: cfn-crawler-jdbc-flights-1
# The name of the database to be created to contain tables
  CFNDatabaseName:
    Type: String
    Default: cfn-database-jdbc-flights-1
# The prefix for all tables crawled and created
  CFNTablePrefixName:
    Type: String
    Default: cfn_jdbc_1_
# The name of the existing connection to the MySQL database
  CFNConnectionName:
    Type: String
    Default: cfn-connection-mysql-flights-1
# The name of the JDBC path (database/schema/table) with wildcard (%) to crawl
  CFNJDBCPath:
    Type: String
    Default: saldev/%
#
#
# Resources section defines metadata for the Data Catalog
Resources:
#Create IAM Role assumed by the crawler. For demonstration, this role is given all
 permissions.
  CFNRoleFlights:
    Type: AWS::IAM::Role
    Properties:
      AssumeRolePolicyDocument:
        Version: "2012-10-17"
        Statement:
          -
            Effect: "Allow"
            Principal:
              Service:
                - "glue.amazonaws.com"
            Action:
              - "sts:AssumeRole"
      Path: "/"
      Policies:
        -
          PolicyName: "root"
          PolicyDocument:
            Version: "2012-10-17"
            Statement:
              -
                Effect: "Allow"
                Action: "*"
                Resource: "*"
 # Create a database to contain tables created by the crawler
  CFNDatabaseFlights:
    Type: AWS::Glue::Database
    Properties:
      CatalogId: !Ref AWS::AccountId
      DatabaseInput:
        Name: !Ref CFNDatabaseName
        Description: "AWS Glue container to hold metadata tables for the flights
 crawler"
 #Create a crawler to crawl the flights data in MySQL database
  CFNCrawlerFlights:
    Type: AWS::Glue::Crawler
    Properties:
      Name: !Ref CFNCrawlerName
      Role: !GetAtt CFNRoleFlights.Arn
      #Classifiers: none, use the default classifier
      Description: AWS Glue crawler to crawl flights data
      #Schedule: none, use default run-on-demand
      DatabaseName: !Ref CFNDatabaseName
      Targets:
        JdbcTargets:
          # JDBC MySQL database with the flights data
          - ConnectionName: !Ref CFNConnectionName
            Path: !Ref CFNJDBCPath
          #Exclusions: none
      TablePrefix: !Ref CFNTablePrefixName
      SchemaChangePolicy:
        UpdateBehavior: "UPDATE_IN_DATABASE"
        DeleteBehavior: "LOG"
   Configuration: "{\"Version\":1.0,\"CrawlerOutput\":{\"Partitions\":
{\"AddOrUpdateBehavior\":\"InheritFromTable\"},\"Tables\":{\"AddOrUpdateBehavior\":
\"MergeNewColumns\"}}}"
```

Sample AWS CloudFormation template for an AWS Glue job for Amazon S3 to Amazon S3

An AWS Glue job in the Data Catalog contains the parameter values that are required to run a script in AWS Glue.

This sample creates a job that reads flight data from an Amazon S3 bucket in `csv` format and writes it to an Amazon S3 Parquet file. The script that is run by this job must already exist. You can generate an ETL script for your environment with the AWS Glue console. When this job is run, an IAM role with the correct permissions must also be supplied.

Common parameter values are shown in the template. For example, `AllocatedCapacity` (DPUs) defaults to 5.

```
---
AWSTemplateFormatVersion: '2010-09-09'
# Sample CFN YAML to demonstrate creating a job using the public flights S3 table in a
 public bucket
#
# Parameters section contains names that are substituted in the Resources section
# These parameters are the names the resources created in the Data Catalog
Parameters:

# The name of the job to be created
  CFNJobName:
    Type: String
    Default: cfn-job-S3-to-S3-2
# The name of the IAM role that the job assumes. It must have access to data, script,
 temporary directory
  CFNIAMRoleName:
    Type: String
    Default: AWSGlueServiceRoleGA
# The S3 path where the script for this job is located
  CFNScriptLocation:
    Type: String
    Default: s3://aws-glue-scripts-123456789012-us-east-1/myid/sal-job-test2
#
#
# Resources section defines metadata for the Data Catalog
Resources:
# Create job to run script which accesses flightscsv table and write to S3 file as
 parquet.
# The script already exists and is called by this job
  CFNJobFlights:
    Type: AWS::Glue::Job
    Properties:
      Role: !Ref CFNIAMRoleName
      #DefaultArguments: JSON object
      # If script written in Scala, then set DefaultArguments={'--job-language';
 'scala', '--class': 'your scala class'}
      #Connections:  No connection needed for S3 to S3 job
      #  ConnectionsList
      #MaxRetries: Double
      Description: Job created with CloudFormation
      #LogUri: String
      Command:
        Name: glueetl
        ScriptLocation: !Ref CFNScriptLocation
             # for access to directories use proper IAM role with permission to buckets
 and folders that begin with "aws-glue-"
             # script uses temp directory from job definition if required (temp
 directory not used S3 to S3)
             # script defines target for output as s3://aws-glue-target/sal
      AllocatedCapacity: 5
      ExecutionProperty:
        MaxConcurrentRuns: 1
      Name: !Ref CFNJobName
```

Sample AWS CloudFormation template for an AWS Glue job for JDBC to Amazon S3

An AWS Glue job in the Data Catalog contains the parameter values that are required to run a script in AWS Glue.

This sample creates a job that reads flight data from a MySQL JDBC database as defined by the connection named `cfn-connection-mysql-flights-1` and writes it to an Amazon S3 Parquet file. The script that is run by this job must already exist. You can generate an ETL script for your environment with the AWS Glue console. When this job is run, an IAM role with the correct permissions must also be supplied.

Common parameter values are shown in the template. For example, `AllocatedCapacity` (DPUs) defaults to 5.

```

---
AWSTemplateFormatVersion: '2010-09-09'
# Sample CFN YAML to demonstrate creating a job using a MySQL JDBC DB with the flights
 data to an S3 file
#
# Parameters section contains names that are substituted in the Resources section
# These parameters are the names the resources created in the Data Catalog
Parameters:

# The name of the job to be created
  CFNJobName:
    Type: String
    Default: cfn-job-JDBC-to-S3-1
```

```
# The name of the IAM role that the job assumes. It must have access to data, script,
 temporary directory
  CFNIAMRoleName:
    Type: String
    Default: AWSGlueServiceRoleGA
# The S3 path where the script for this job is located
  CFNScriptLocation:
    Type: String
    Default: s3://aws-glue-scripts-123456789012-us-east-1/myid/sal-job-dec4a
# The name of the connection used for JDBC data source
  CFNConnectionName:
    Type: String
    Default: cfn-connection-mysql-flights-1
#
#
# Resources section defines metadata for the Data Catalog
Resources:
# Create job to run script which accesses JDBC flights table via a connection and write
 to S3 file as parquet.
# The script already exists and is called by this job
  CFNJobFlights:
    Type: AWS::Glue::Job
    Properties:
      Role: !Ref CFNIAMRoleName
      #DefaultArguments: JSON object
      # For example, if required by script, set temporary directory as
 DefaultArguments={'--TempDir'; 's3://aws-glue-temporary-xyc/sal'}
      Connections:
        Connections:
        - !Ref CFNConnectionName
      #MaxRetries: Double
      Description: Job created with CloudFormation using existing script
      #LogUri: String
      Command:
        Name: glueetl
        ScriptLocation: !Ref CFNScriptLocation
             # for access to directories use proper IAM role with permission to buckets
 and folders that begin with "aws-glue-"
             # if required, script defines temp directory as argument TempDir and used
 in script like redshift_tmp_dir = args["TempDir"]
             # script defines target for output as s3://aws-glue-target/sal
      AllocatedCapacity: 5
      ExecutionProperty:
        MaxConcurrentRuns: 1
      Name: !Ref CFNJobName
```

Sample AWS CloudFormation template for an AWS Glue on-demand trigger

An AWS Glue trigger in the Data Catalog contains the parameter values that are required to start a job run when the trigger fires. An on-demand trigger fires when you enable it.

This sample creates an on-demand trigger that starts one job named `cfn-job-S3-to-S3-1`.

```
---
AWSTemplateFormatVersion: '2010-09-09'
# Sample CFN YAML to demonstrate creating an on-demand trigger
#
# Parameters section contains names that are substituted in the Resources section
# These parameters are the names the resources created in the Data Catalog
Parameters:
  # The existing job to be started by this trigger
  CFNJobName:
    Type: String
    Default: cfn-job-S3-to-S3-1
  # The name of the trigger to be created
  CFNTriggerName:
    Type: String
    Default: cfn-trigger-ondemand-flights-1
#
# Resources section defines metadata for the Data Catalog
# Sample CFN YAML to demonstrate creating an on-demand trigger for a job
Resources:
# Create trigger to run an existing job (CFNJobName) on an on-demand schedule.
  CFNTriggerSample:
    Type: AWS::Glue::Trigger
    Properties:
      Name:
        Ref: CFNTriggerName
      Description: Trigger created with CloudFormation
      Type: ON_DEMAND
      Actions:
        - JobName: !Ref CFNJobName
        # Arguments: JSON object
      #Schedule:
```

```
      #Predicate:
```

Sample AWS CloudFormation template for an AWS Glue scheduled trigger

An AWS Glue trigger in the Data Catalog contains the parameter values that are required to start a job run when the trigger fires. A scheduled trigger fires when it is enabled and the cron timer pops.

This sample creates a scheduled trigger that starts one job named `cfn-job-S3-to-S3-1`. The timer is a cron expression to run the job every 10 minutes on weekdays.

```
---
AWSTemplateFormatVersion: '2010-09-09'
# Sample CFN YAML to demonstrate creating a scheduled trigger
#
# Parameters section contains names that are substituted in the Resources section
# These parameters are the names the resources created in the Data Catalog
Parameters:
  # The existing job to be started by this trigger
  CFNJobName:
    Type: String
    Default: cfn-job-S3-to-S3-1
  # The name of the trigger to be created
  CFNTriggerName:
    Type: String
    Default: cfn-trigger-scheduled-flights-1
#
# Resources section defines metadata for the Data Catalog
# Sample CFN YAML to demonstrate creating a scheduled trigger for a job
#
Resources:
# Create trigger to run an existing job (CFNJobName) on a cron schedule.
  TriggerSample1CFN:
    Type: AWS::Glue::Trigger
    Properties:
      Name:
        Ref: CFNTriggerName
      Description: Trigger created with CloudFormation
      Type: SCHEDULED
      Actions:
        - JobName: !Ref CFNJobName
        # Arguments: JSON object
      # # Run the trigger every 10 minutes on Monday to Friday
      Schedule: cron(0/10 * ? * MON-FRI *)
      #Predicate:
```

Sample AWS CloudFormation template for an AWS Glue conditional trigger

An AWS Glue trigger in the Data Catalog contains the parameter values that are required to start a job run when the trigger fires. A conditional trigger fires when it is enabled and its conditions are met, such as a job completing successfully.

This sample creates a conditional trigger that starts one job named `cfn-job-S3-to-S3-1`. This job starts when the job named `cfn-job-S3-to-S3-2` completes successfully.

```
---
AWSTemplateFormatVersion: '2010-09-09'
# Sample CFN YAML to demonstrate creating a conditional trigger for a job, which starts
 when another job completes
#
# Parameters section contains names that are substituted in the Resources section
# These parameters are the names the resources created in the Data Catalog
Parameters:
  # The existing job to be started by this trigger
  CFNJobName:
    Type: String
    Default: cfn-job-S3-to-S3-1
  # The existing job that when it finishes causes trigger to fire
  CFNJobName2:
    Type: String
    Default: cfn-job-S3-to-S3-2
  # The name of the trigger to be created
  CFNTriggerName:
    Type: String
    Default: cfn-trigger-conditional-1
#
Resources:
# Create trigger to run an existing job (CFNJobName) when another job completes
 (CFNJobName2).
  CFNTriggerSample:
    Type: AWS::Glue::Trigger
    Properties:
      Name:
        Ref: CFNTriggerName
      Description: Trigger created with CloudFormation
      Type: CONDITIONAL
      Actions:
        - JobName: !Ref CFNJobName
        # Arguments: JSON object
      #Schedule: none
      Predicate:
        #Value for Logical is required if more than 1 job listed in Conditions
        Logical: AND
        Conditions:
          - LogicalOperator: EQUALS
            JobName: !Ref CFNJobName2
            State: SUCCEEDED
```

Sample AWS CloudFormation template for an AWS Glue development endpoint

An AWS Glue machine learning transform is a custom transform to cleanse your data. There is currently one available transform named FindMatches. The FindMatches transform enables you to identify duplicate or matching records in your dataset, even when the records do not have a common unique identifier and no fields match exactly.

This sample creates a machine learning transform. For more information about the parameters that you need to create a machine learning transform, see Record matching with AWS Lake Formation FindMatches.

```

---
AWSTemplateFormatVersion: '2010-09-09'
# Sample CFN YAML to demonstrate creating a machine learning transform
#
# Resources section defines metadata for the machine learning transform
Resources:
  MyMLTransform:
    Type: "AWS::Glue::MLTransform"
    Condition: "isGlueMLGARegion"
    Properties:
      Name: !Sub "MyTransform"
      Description: "The bestest transform ever"
      Role: !ImportValue MyMLTransformUserRole
      GlueVersion: "1.0"
      WorkerType: "Standard"
      NumberOfWorkers: 5
      Timeout: 120
      MaxRetries: 1
      InputRecordTables:
        GlueTables:
          - DatabaseName: !ImportValue MyMLTransformDatabase
            TableName: !ImportValue MyMLTransformTable
      TransformParameters:
        TransformType: "FIND_MATCHES"
        FindMatchesParameters:
          PrimaryKeyColumnName: "testcolumn"
          PrecisionRecallTradeoff: 0.5
          AccuracyCostTradeoff: 0.5
          EnforceProvidedLabels: True
      Tags:
        key1: "value1"
        key2: "value2"
      TransformEncryption:
        TaskRunSecurityConfigurationName: !ImportValue
 MyMLTransformSecurityConfiguration
        MLUserDataEncryption:
          MLUserDataEncryptionMode: "SSE-KMS"
          KmsKeyId: !ImportValue MyMLTransformEncryptionKey
```

Sample AWS CloudFormation template for an AWS Glue Data Quality ruleset

An AWS Glue Data Quality ruleset contains rules that can be evaluated on a table within the Data Catalog. Once the ruleset is placed on your targeted table you can go into the Data Catalog and run an evaluation which runs your data against those rules within the ruleset. These rules can vary from evaluating the row count to evaluating referential integrity on your data.

The following sample is a CloudFormation template which creates a ruleset with a variety of rules on the specified target table.

```
AWSTemplateFormatVersion: '2010-09-09'
```

```
# Sample CFN YAML to demonstrate creating a DataQualityRuleset
#
# Parameters section contains names that are substituted in the Resources section
# These parameters are the names the resources created in the Data Catalog
Parameters:

  # The name of the ruleset to be created
  RulesetName:
    Type: String
    Default: "CFNRulesetName"
  RulesetDescription:
    Type: String
    Default: "CFN DataQualityRuleset"
  # Rules that will be associated with this ruleset
  Rules:
    Type: String
    Default: 'Rules = [
        RowCount > 100,
        IsUnique "id",
        IsComplete "nametype"
        ]'
  # Name of database and table within Data Catalog which the ruleset will
  # be applied too
  DatabaseName:
    Type: String
    Default: "ExampleDatabaseName"
  TableName:
    Type: String
    Default: "ExampleTableName"

# Resources section defines metadata for the Data Catalog
Resources:
  # Creates a Data Quality ruleset under specified rules
  DQRuleset:
    Type: AWS::Glue::DataQualityRuleset
    Properties:
      Name: !Ref RulesetName
      Description: !Ref RulesetDescription
      # The String within rules must be formatted in DQDL, a language
      # used specifically to make rules
      Ruleset: !Ref Rules
      # The targeted table must exist within Data Catalog alongside
      # the correct database
      TargetTable:
        DatabaseName: !Ref DatabaseName
        TableName: !Ref TableName
```

Sample AWS CloudFormation template for an AWS Glue Data Quality ruleset with EventBridge scheduler

An AWS Glue Data Quality ruleset contains rules that can be evaluated on a table within the Data Catalog. Once the ruleset is placed on your targeted table you can go into the Data Catalog and run an evaluation which runs your data against those rules within the ruleset. Instead of having to manually go into the Data Catalog to evaluate the ruleset, you can also add an EventBridge Scheduler within our CloudFormation template to schedule these ruleset evaluations for you on a timed interval.

The following sample is a CloudFormation template which creates a Data Quality ruleset and a EventBridge Scheduler to evaluate the aforementioned ruleset every five minutes.

```
AWSTemplateFormatVersion: '2010-09-09'
# Sample CFN YAML to demonstrate creating a DataQualityRuleset
#
# Parameters section contains names that are substituted in the Resources section
# These parameters are the names the resources created in the Data Catalog
Parameters:

  # The name of the ruleset to be created
  RulesetName:
    Type: String
    Default: "CFNRulesetName"
  # Rules that will be associated with this Ruleset
  Rules:
    Type: String
    Default: 'Rules = [
        RowCount > 100,
        IsUnique "id",
        IsComplete "nametype"
        ]'
  # The name of the Schedule to be created
  ScheduleName:
    Type: String
    Default: "ScheduleDQRulsetEvaluation"
  # This expression determines the rate at which the Schedule will evaluate
  # your data using the above ruleset
  ScheduleRate:
    Type: String
    Default: "rate(5 minutes)"
  # The Request that being sent must match the details of the Data Quality Ruleset
  ScheduleRequest:
    Type: String
    Default: '
        { "DataSource": { "GlueTable": { "DatabaseName": "ExampleDatabaseName",
         "TableName": "ExampleTableName" } },
         "Role": "role/AWSGlueServiceRoleDefault",
          "RulesetNames": [ ""CFNRulesetName"" ] }
        '

# Resources section defines metadata for the Data Catalog
Resources:
  # Creates a Data Quality ruleset under specified rules
  DQRuleset:
    Type: AWS::Glue::DataQualityRuleset
    Properties:
      Name: !Ref RulesetName
      Description: "CFN DataQualityRuleset"
      # The String within rules must be formatted in DQDL, a language
      # used specifically to make rules
      Ruleset: !Ref Rules
      # The targeted table must exist within Data Catalog alongside
      # the correct database
      TargetTable:
        DatabaseName: "ExampleDatabaseName"
        TableName: "ExampleTableName"
  # Create a Scheduler to schedule evaluation runs on the above ruleset
  ScheduleDQEval:
    Type: AWS::Scheduler::Schedule
    Properties:
      Name: !Ref ScheduleName
      Description: "Schedule DataQualityRuleset Evaluations"
      FlexibleTimeWindow:
        Mode: "OFF"
      ScheduleExpression: !Ref ScheduleRate
      ScheduleExpressionTimezone: "America/New_York"
      State: "ENABLED"
      Target:
        # The ARN is the API that will be run, since we want to evaluate our ruleset
        # we want this specific ARN
        Arn: "arn:aws:scheduler:::aws-sdk:glue:startDataQualityRulesetEvaluationRun"
        # Your RoleArn must have approval to schedule
        RoleArn: "arn:aws:iam::123456789012:role/AWSGlueServiceRoleDefault"
        # This is the Request that is being sent to the Arn
        Input: '
        { "DataSource": { "GlueTable": { "DatabaseName": "sampledb", "TableName":
 "meteorite" } },
         "Role": "role/AWSGlueServiceRoleDefault",
          "RulesetNames": [ "TestCFN" ] }
        '
```

Sample AWS CloudFormation template for an AWS Glue development endpoint

An AWS Glue development endpoint is an environment that you can use to develop and test your AWS Glue scripts.

This sample creates a development endpoint with the minimal network parameter values required to successfully create it. For more information about the parameters that you need to set up a development endpoint, see Setting up networking for development for AWS Glue.

You provide an existing IAM role ARN (Amazon Resource Name) to create the development endpoint. Supply a valid RSA public key and keep the corresponding private key available if you plan to create a notebook server on the development endpoint.

> **Note**
>
> For any notebook server that you create that is associated with a development endpoint, you manage it. Therefore, if you delete the development endpoint, to delete the notebook server, you must delete the AWS CloudFormation stack on the AWS CloudFormation console.

```

---
AWSTemplateFormatVersion: '2010-09-09'
# Sample CFN YAML to demonstrate creating a development endpoint
#
# Parameters section contains names that are substituted in the Resources section
# These parameters are the names the resources created in the Data Catalog
```

```
Parameters:

# The name of the crawler to be created
  CFNEndpointName:
    Type: String
    Default: cfn-devendpoint-1
  CFNIAMRoleArn:
    Type: String
    Default: arn:aws:iam::123456789012/role/AWSGlueServiceRoleGA
#
#
# Resources section defines metadata for the Data Catalog
Resources:
  CFNDevEndpoint:
    Type: AWS::Glue::DevEndpoint
    Properties:
      EndpointName: !Ref CFNEndpointName
      #ExtraJarsS3Path: String
      #ExtraPythonLibsS3Path: String
      NumberOfNodes: 5
      PublicKey: ssh-rsa public.....key myuserid-key
      RoleArn: !Ref CFNIAMRoleArn
      SecurityGroupIds:
        - sg-64986c0b
      SubnetId: subnet-c67cccac
```

AWS Glue programming guide

A script contains the code that extracts data from sources, transforms it, and loads it into targets. AWS Glue runs a script when it starts a job.

AWS Glue ETL scripts are coded in Python or Scala. While all job types can be written in Python, AWS Glue for Spark jobs can be written in Scala as well. When you automatically generate the source code logic for your job in AWS Glue Studio, a script is created. You can edit this script, or you can provide your own script to process your ETL work.

Providing your own custom scripts

Scripts perform the extract, transform, and load (ETL) work in AWS Glue. A script is created when you automatically generate the source code logic for a job. You can either edit this generated script, or you can provide your own custom script.

To provide your own custom script in AWS Glue, follow these general steps:

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. Choose the **ETL Jobs** tab, and then view the **Create job** section. Choose a **script editor** option.
3. Under **This job runs**, choose one of the following:
    - **Create a new script with boilerplate code**
    - **Upload and edit an existing script**
4. On the **Job details** page, choose the **IAM role** that is required for your custom script to run. For more information, see Identity and access management for AWS Glue.
5. Choose any connections that your script references. These objects are needed to connect to the necessary JDBC data stores.

    An elastic network interface is a virtual network interface that you can attach to an instance in a virtual private cloud (VPC). Choose the elastic network interface that is required to connect to the data store that's used in the script.
6. Provide additional configuration, including parameters, specific to your job type. For more information about configuration for your job type, see the *Building visual ETL jobs* section.

7. On the **Script** tab, paste or write your custom script.

Use the content in this section to guide the process of writing your custom script.

For more information about adding jobs in AWS Glue, see *Building visual ETL jobs*.

For step-by-step guidance, see the **Add job** tutorial in the AWS Glue console.

Programming Spark scripts

AWS Glue makes it easy to write or autogenerate extract, transform, and load (ETL) scripts, in addition to testing and running them. This section describes the extensions to Apache Spark that AWS Glue has introduced, and provides examples of how to code and run ETL scripts in Python and Scala.

> ⚠ **Important**
>
> Different versions of AWS Glue support different versions of Apache Spark. Your custom script must be compatible with the supported Apache Spark version. For information about AWS Glue versions, see the Glue version job property.

Topics

- Tutorial: Writing an AWS Glue for Spark script
- Program AWS Glue ETL scripts in PySpark
- Programming AWS Glue ETL scripts in Scala
- Features and optimizations for programming AWS Glue for Spark ETL scripts

Tutorial: Writing an AWS Glue for Spark script

This tutorial introduces you to the process of writing AWS Glue scripts. You can run scripts on a schedule with jobs, or interactively with interactive sessions. For more information about jobs, see *Building visual ETL jobs*. For more information about interactive sessions, see the section called “Overview of AWS Glue interactive sessions”.

The AWS Glue Studio visual editor offers a graphical, no-code interface for building AWS Glue jobs. AWS Glue scripts back visual jobs. They give you access to the expanded set of tools available to work with Apache Spark programs. You can access native Spark APIs, as well as AWS Glue libraries that facilitate extract, transform, and load (ETL) workflows from within an AWS Glue script.

In this tutorial, you extract, transform, and load a dataset of parking tickets. The script that does this work is identical in form and function to the one generated in Making ETL easier with AWS Glue Studio on the AWS Big Data Blog, which introduces the AWS Glue Studio visual editor. By running this script in a job, you can compare it to visual jobs and see how AWS Glue ETL scripts work. This prepares you to use additional functionalities that aren't yet available in visual jobs.

You use the Python language and libraries in this tutorial. Similar functionality is available in Scala. After going through this tutorial, you should be able to generate and inspect a sample Scala script to understand how to perform the Scala AWS Glue ETL script writing process.

Prerequisites

This tutorial has the following prerequisites:

- The same prerequisites as the AWS Glue Studio blog post, which instructs you to run a AWS CloudFormation template.

  This template uses the AWS Glue Data Catalog to manage the parking ticket dataset available in `s3://aws-bigdata-blog/artifacts/gluestudio/`. It creates the following resources which will be referenced:
- **AWS Glue StudioRole** – IAM role to run AWS Gluejobs
- **AWS Glue StudioAmazon S3Bucket** – Name of the Amazon S3 bucket to store blog-related files
- **AWS Glue StudioTicketsYYZDB** – AWS Glue Data Catalog database
- **AWS Glue StudioTableTickets** – Data Catalog table to use as a source
- **AWS Glue StudioTableTrials** – Data Catalog table to use as a source
- **AWS Glue StudioParkingTicketCount** – Data Catalog table to use as the destination
- The script generated in the AWS Glue Studio blog post. If the blog post changes, the script is also available in the following text.

Generate a sample script

You can use the AWS Glue Studio visual editor as a powerful code generation tool to create a scaffold for the script you want to write. You will use this tool to create a sample script.

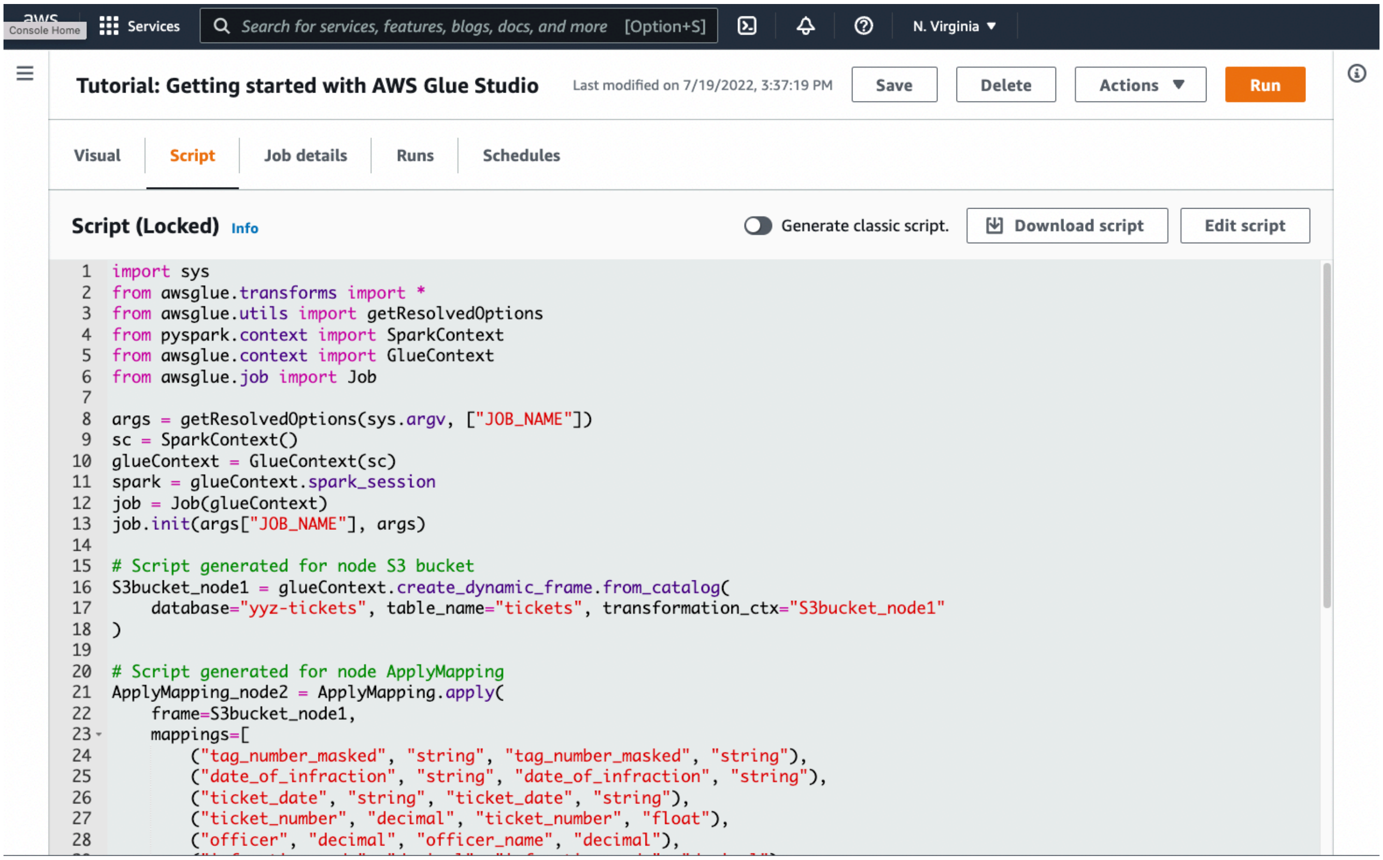
If you want to skip these steps, the script is provided.

Tutorial sample script

```
import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job

args = getResolvedOptions(sys.argv, ["JOB_NAME"])
sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
job.init(args["JOB_NAME"], args)

# Script generated for node S3 bucket
S3bucket_node1 = glueContext.create_dynamic_frame.from_catalog(
    database="yyz-tickets", table_name="tickets", transformation_ctx="S3bucket_node1"
)

# Script generated for node ApplyMapping
ApplyMapping_node2 = ApplyMapping.apply(
    frame=S3bucket_node1,
    mappings=[
        ("tag_number_masked", "string", "tag_number_masked", "string"),
        ("date_of_infraction", "string", "date_of_infraction", "string"),
        ("ticket_date", "string", "ticket_date", "string"),
        ("ticket_number", "decimal", "ticket_number", "float"),
        ("officer", "decimal", "officer_name", "decimal"),
        ("infraction_code", "decimal", "infraction_code", "decimal"),
        ("infraction_description", "string", "infraction_description", "string"),
        ("set_fine_amount", "decimal", "set_fine_amount", "float"),
        ("time_of_infraction", "decimal", "time_of_infraction", "decimal"),
    ],
    transformation_ctx="ApplyMapping_node2",
)

# Script generated for node S3 bucket
S3bucket_node3 = glueContext.write_dynamic_frame.from_options(
    frame=ApplyMapping_node2,
    connection_type="s3",
    format="glueparquet",
    connection_options={"path": "s3://DOC-EXAMPLE-BUCKET", "partitionKeys": []},
    format_options={"compression": "gzip"},
    transformation_ctx="S3bucket_node3",
)

job.commit()
```

To generate a sample script

1. Complete the AWS Glue Studio tutorial. To complete this tutorial, see [Creating a job in AWS Glue Studio from an example job](#).
2. Navigate to the **Script** tab on the job page, as shown in the following screenshot:


aws
Console Home
Services
Search for services, features, blogs, docs, and more [Option+S]
N. Virginia ▼
Tutorial: Getting started with AWS Glue Studio
Last modified on 7/19/2022, 3:37:19 PM
Save
Delete
Actions ▼
Run
Visual
Script
Job details
Runs
Schedules
Script (Locked)
Info
Generate classic script.
Download script
Edit script

```
import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job

args = getResolvedOptions(sys.argv, ["JOB_NAME"])
sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
job.init(args["JOB_NAME"], args)

# Script generated for node S3 bucket
S3bucket_node1 = glueContext.create_dynamic_frame.from_catalog(
    database="yyz-tickets", table_name="tickets", transformation_ctx="S3bucket_node1"
)

# Script generated for node ApplyMapping
ApplyMapping_node2 = ApplyMapping.apply(
    frame=S3bucket_node1,
    mappings=[
        ("tag_number_masked", "string", "tag_number_masked", "string"),
        ("date_of_infraction", "string", "date_of_infraction", "string"),
        ("ticket_date", "string", "ticket_date", "string"),
        ("ticket_number", "decimal", "ticket_number", "float"),
        ("officer", "decimal", "officer_name", "decimal"),
```

3. Copy the complete contents of the **Script** tab. By setting the script language in **Job details**, you can switch back and forth between generating Python or Scala code.

Step 1. Create a job and paste your script

In this step, you create an AWS Glue job in the AWS Management Console. This sets up a configuration that allows AWS Glue to run your script. It also creates a place for you to store and edit your script.

To create a job

1. In the AWS Management Console, navigate to the AWS Glue landing page.
2. In the side navigation pane, choose **Jobs**.
3. Choose **Spark script editor** in **Create job**, and then choose **Create**.
4. **Optional** - Paste the full text of your script into the **Script** pane. Alternatively, you can follow along with the tutorial.

Step 2. Import AWS Glue libraries

You need to set your script up to interact with code and configuration that are defined outside of the script. This work is done behind the scenes in AWS Glue Studio.

In this step, you perform the following actions.

- Import and initialize a `GlueContext` object. This is the most important import, from the script writing perspective. This exposes standard methods for defining source and target datasets, which is the starting point for any ETL script. To learn more about the `GlueContext` class, see GlueContext class.
- Initialize a `SparkContext` and `SparkSession`. These allow you to configure the Spark engine available inside the AWS Glue job. You won't need to use them directly within introductory AWS Glue scripts.
- Call `getResolvedOptions` to prepare your job arguments for use within the script. For more information about resolving job parameters, see the section called "getResolvedOptions".
- Initialize a `Job`. The `Job` object sets configuration and tracks the state of various optional AWS Glue features. Your script can run without a `Job` object, but the best practice is to initialize it so that you don't encounter confusion if those features are later integrated.

  One of these features is job bookmarks, which you can optionally configure in this tutorial. You can learn about job bookmarks in the following section, the section called "Optional - Enable job bookmarks".

In this procedure, you write the following code. This code is a portion of the generated sample script.

```
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job

args = getResolvedOptions(sys.argv, ["JOB_NAME"])
sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
job.init(args["JOB_NAME"], args)
```

To import AWS Glue libraries

- Copy this section of code and paste it into the **Script** editor.

> **Note**
>
> You might consider copying code to be a bad engineering practice. In this tutorial, we suggest this to encourage you to consistently name your core variables across all AWS Glue ETL scripts.

Step 3. Extract data from a source

In any ETL process, you first need to define a source dataset that you want to change. In the AWS Glue Studio visual editor, you provide this information by creating a **Source** node.

In this step, you provide the `create_dynamic_frame.from_catalog` method a `database` and `table_name` to extract data from a source configured in the AWS Glue Data Catalog.

In the previous step, you initialized a `GlueContext` object. You use this object to find methods that are used to configure sources, such as `create_dynamic_frame.from_catalog`.

In this procedure, you write the following code using `create_dynamic_frame.from_catalog`. This code is a portion of the generated sample script.

```
S3bucket_node1 = glueContext.create_dynamic_frame.from_catalog(
    database="yyz-tickets", table_name="tickets", transformation_ctx="S3bucket_node1"
    )
```

To extract data from a source

1. Examine the documentation to find a method on `GlueContext` to extract data from a source defined in the AWS Glue Data Catalog. These methods are documented in the section called “GlueContext”. Choose the create_dynamic_frame.from_catalog method. Call this method on `glueContext`.
2. Examine the documentation for `create_dynamic_frame.from_catalog`. This method requires `database` and `table_name` parameters. Provide the necessary parameters to `create_dynamic_frame.from_catalog`.

   The AWS Glue Data Catalog stores information about the location and format of your source data, and was set up in the prerequisite section. You don't have to directly provide your script with that information.
3. **Optional** – Provide the `transformation_ctx` parameter to the method in order to support job bookmarks. You can learn about job bookmarks in the following section, the section called “Optional - Enable job bookmarks”.

> **Note**
>
> **Common methods for extracting data**
>
> the section called “create_dynamic_frame_from_catalog” is used to connect to tables in the AWS Glue Data Catalog.
>
> If you need to directly provide your job with configuration that describes the structure and location of your source, see the the section called “create_dynamic_frame_from_options” method. You will need to provide more detailed parameters describing your data than when using `create_dynamic_frame.from_catalog`.
>
> Refer to the supplemental documentation about `format_options` and `connection_parameters` to identify your required parameters. For an explanation of how to provide your script information about your source data format, see the section called “Data format options”. For an explanation of how to provide your script information about your source data location, see the section called “Connection parameters”.
>
> If you're reading information from a streaming source, you provide your job with source information through the the section called “create_data_frame_from_catalog” or the section called “create_data_frame_from_options” methods. Note that these methods return Apache Spark `DataFrames`.

Our generated code calls `create_dynamic_frame.from_catalog` while the reference documentation refers to `create_dynamic_frame_from_catalog`. These methods ultimately call the same code, and are included so you can write cleaner code. You can verify this by viewing the source for our Python wrapper, available at aws-glue-libs.

Step 4. Transform data with AWS Glue

After extracting source data in an ETL process, you need to describe how you want to change your data. You provide this information by creating a **Transform** node in the AWS Glue Studio visual editor.

In this step, you provide the `ApplyMapping` method with a map of current and desired field names and types to transform your `DynamicFrame`.

You perform the following transformations.

- Drop the four `location` and `province` keys.
- Change the name of `officer` to `officer_name`.
- Change the type of `ticket_number` and `set_fine_amount` to `float`.

`create_dynamic_frame.from_catalog` provides you with a `DynamicFrame` object. A `DynamicFrame` represents a dataset in AWS Glue. AWS Glue transforms are operations that change `DynamicFrames`.

Note

What is a `DynamicFrame`?
A `DynamicFrame` is an abstraction that allows you to connect a dataset with a description of the names and types of entries in the data. In Apache Spark, a similar abstraction exists called a DataFrame. For an explanation of DataFrames, see Spark SQL Guide.
With `DynamicFrames`, you can describe dataset schemas dynamically. Consider a dataset with a price column, where some entries store price as a string, and others store price as a double. AWS Glue computes a schema on-the-fly—it creates a self-describing record for each row.
Inconsistent fields (like price) are explicitly represented with a type (`ChoiceType`) in the schema for the frame. You can address your inconsistent fields by dropping them with `DropFields` or resolving them with `ResolveChoice`. These are transforms that are

available on the `DynamicFrame`. You can then write your data back to your data lake with `writeDynamicFrame`.

You can call many of the same transforms from methods on the `DynamicFrame` class, which can lead to more readable scripts. For more information about `DynamicFrame`, see [the section called “DynamicFrame”](#).

In this procedure, you write the following code using `ApplyMapping`. This code is a portion of the generated sample script.

```
ApplyMapping_node2 = ApplyMapping.apply(
    frame=S3bucket_node1,
    mappings=[
        ("tag_number_masked", "string", "tag_number_masked", "string"),
        ("date_of_infraction", "string", "date_of_infraction", "string"),
        ("ticket_date", "string", "ticket_date", "string"),
        ("ticket_number", "decimal", "ticket_number", "float"),
        ("officer", "decimal", "officer_name", "decimal"),
        ("infraction_code", "decimal", "infraction_code", "decimal"),
        ("infraction_description", "string", "infraction_description", "string"),
        ("set_fine_amount", "decimal", "set_fine_amount", "float"),
        ("time_of_infraction", "decimal", "time_of_infraction", "decimal"),
    ],
    transformation_ctx="ApplyMapping_node2",
)
```

To transform data with AWS Glue

1. Examine the documentation to identify a transform to change and drop fields. For details, see [the section called “GlueTransform”](#). Choose the `ApplyMapping` transform. For more information about `ApplyMapping`, see [the section called “ApplyMapping”](#). Call `apply` on the `ApplyMapping` transform object.

   > **Note**
   >
   > What is `ApplyMapping`?
   >
   > `ApplyMapping` takes a `DynamicFrame` and transforms it. It takes a list of tuples that represent transformations on fields—a "mapping". The first two tuple elements, a field name and type, are used to identify a field in the frame. The second two parameters are also a field name and type.

ApplyMapping converts the source field to the target name and type in a new `DynamicFrame`, which it returns. Fields that aren't provided are dropped in the return value.
Rather than calling `apply`, you can call the same transform with the `apply_mapping` method on the `DynamicFrame` to create more fluent, readable code. For more information, see the section called “apply_mapping”.

2. Examine the documentation for `ApplyMapping` to identify required parameters. See the section called “ApplyMapping”. You will find that this method requires `frame` and `mappings` parameters. Provide the necessary parameters to `ApplyMapping`.
3. **Optional** – Provide `transformation_ctx` to the method to support job bookmarks. You can learn about job bookmarks in the following section, the section called “Optional - Enable job bookmarks”.

Note

Apache Spark functionality
We provide transforms to streamline ETL workflows within your job. You also have access to the libraries that are available in a Spark program in your job, built for more general purposes. In order to use them, you convert between `DynamicFrame` and `DataFrame`. You can create a `DataFrame` with the section called “toDF”. Then, you can use methods available on the DataFrame to transform your dataset. For more information on these methods, see DataFrame. You can then convert backwards with the section called “fromDF” to use AWS Glue operations for loading your frame to a target.

Step 5. Load data into a target

After you transform your data, you typically store the transformed data in a different place from the source. You perform this operation by creating a **target** node in the AWS Glue Studio visual editor.

In this step, you provide the `write_dynamic_frame.from_options` method a `connection_type`, `connection_options`, `format`, and `format_options` to load data into a target bucket in Amazon S3.

In Step 1, you initialized a `GlueContext` object. In AWS Glue, this is where you will find methods that are used to configure targets, much like sources.

In this procedure, you write the following code using `write_dynamic_frame.from_options`. This code is a portion of the generated sample script.

```
S3bucket_node3 = glueContext.write_dynamic_frame.from_options(
    frame=ApplyMapping_node2,
    connection_type="s3",
    format="glueparquet",
    connection_options={"path": "s3://amzn-s3-demo-bucket", "partitionKeys": []},
    format_options={"compression": "gzip"},
    transformation_ctx="S3bucket_node3",
    )
```

To load data into a target

1. Examine the documentation to find a method to load data into a target Amazon S3 bucket. These methods are documented in the section called “GlueContext”. Choose the the section called “write_dynamic_frame_from_options” method. Call this method on `glueContext`.

   > **Note**
   >
   > **Common methods for loading data**
   > `write_dynamic_frame.from_options` is the most common method used to load data. It supports all targets that are available in AWS Glue.
   > If you're writing to a JDBC target defined in an AWS Glue connection, use the the section called “write_dynamic_frame_from_jdbc_conf” method. AWS Glue connections store information about how to connect to a data source. This removes the need to provide that information in `connection_options`. However, you still need to use `connection_options` to provide `dbtable`.
   > `write_dynamic_frame.from_catalog` is not a common method for loading data. This method updates the AWS Glue Data Catalog without updating the underlying dataset, and is used in combination with other processes that change the underlying dataset. For more information, see the section called “Updating the schema and adding new partitions”.

2. Examine the documentation for the section called “write_dynamic_frame_from_options”. This method requires `frame`, `connection_type`, `format`, `connection_options`, and `format_options`. Call this method on `glueContext`.

    a. Refer to the supplemental documentation about `format_options` and `format` to identify the parameters you need. For an explanation of data formats, see the section called “Data format options”.

    b. Refer to the supplemental documentation about `connection_type` and `connection_options` to identify the parameters you need. For an explanation of connections, see the section called “Connection parameters”.

    c. Provide the necessary parameters to `write_dynamic_frame.from_options`. This method has a similar configuration to `create_dynamic_frame.from_options`.

3. **Optional** – Provide `transformation_ctx` to `write_dynamic_frame.from_options` to support job bookmarks. You can learn about job bookmarks in the following section, the section called “Optional - Enable job bookmarks”.

Step 6. Commit the Job object

You initialized a `Job` object in Step 1. You may need to manually conclude its lifecycle at the end of your script if certain optional features need this to function properly, such as when using Job Bookmarks. This work is done behind the scenes in AWS Glue Studio.

In this step, call the `commit` method on the `Job` object.

In this procedure, you write the following code. This code is a portion of the generated sample script.

```
job.commit()
```

To commit the Job object

1. If you have not yet done this, perform the optional steps outlined in previous sections to include `transformation_ctx`.
2. Call `commit`.

Optional - Enable job bookmarks

In every prior step, you have been instructed to set `transformation_ctx` parameters. This is related to a feature called job bookmarks.

With job bookmarks, you can save time and money with jobs that run on a recurring basis, against datasets where previous work can easily be tracked. Job bookmarks track the progress of an AWS Glue transform across a dataset from previous runs. By tracking where previous runs ended, AWS Glue can limit its work to rows it hasn't processed before. For more information about job bookmarks, see the section called “Tracking processed data using job bookmarks”.

To enable job bookmarks, first add the `transformation_ctx` statements into our provided functions, as described in the previous examples. Job bookmark state is persisted across runs. `transformation_ctx` parameters are keys used to access that state. On their own, these statements will do nothing. You also need to activate the feature in the configuration for your job.

In this procedure, you enable job bookmarks using the AWS Management Console.

To enable job bookmarks

1. Navigate to the **Job details** section of your corresponding job.
2. Set **Job bookmark** to **Enable**.

Step 7. Run your code as a job

In this step, you run your job to verify that you successfully completed this tutorial. This is done with the click of a button, as in the AWS Glue Studio visual editor.

To run your code as a job

1. Choose **Untitled job** on the title bar to edit and set your job name.
2. Navigate to the **Job details** tab. Assign your job an **IAM Role**. You can use the one created by the AWS CloudFormation template in the prerequisites for the AWS Glue Studio tutorial. If you have completed that tutorial, it should be available as `AWS Glue StudioRole`.
3. Choose **Save** to save your script.
4. Choose **Run** to run your job.
5. Navigate to the **Runs** tab to verify that your job completes.

6. Navigate to *`amzn-s3-demo-bucket`*, the target for `write_dynamic_frame.from_options`. Confirm that the output matches your expectations.

For more information about configuring and managing jobs, see the section called “Providing your own custom scripts”.

More information

Apache Spark libraries and methods are available in AWS Glue scripts. You can look at the Spark documentation to understand what you can do with those included libraries. For more information, see the examples section of the Spark source repository.

AWS Glue 2.0+ includes several common Python libraries by default. There are also mechanisms for loading your own dependencies into an AWS Glue job in a Scala or Python environment. For information about Python dependencies, see the section called “Python libraries”.

For more examples of how to use AWS Glue features in Python, see the section called “Python samples”. Scala and Python jobs have feature parity, so our Python examples should give you some thoughts about how to perform similar work in Scala.

Program AWS Glue ETL scripts in PySpark

You can find Python code examples and utilities for AWS Glue in the AWS Glue samples repository on the GitHub website.

Using Python with AWS Glue

AWS Glue supports an extension of the PySpark Python dialect for scripting extract, transform, and load (ETL) jobs. This section describes how to use Python in ETL scripts and with the AWS Glue API.

- Setting up to use Python with AWS Glue
- Calling AWS Glue APIs in Python
- Using Python libraries with AWS Glue
- AWS Glue Python code samples

AWS Glue PySpark extensions

AWS Glue has created the following extensions to the PySpark Python dialect.

- Accessing parameters using `getResolvedOptions`
- PySpark extension types
- DynamicFrame class
- DynamicFrameCollection class
- DynamicFrameWriter class
- DynamicFrameReader class
- GlueContext class

AWS Glue PySpark transforms

AWS Glue has created the following transform Classes to use in PySpark ETL operations.

- GlueTransform base class
- ApplyMapping class
- DropFields class
- DropNullFields class
- ErrorsAsDynamicFrame class
- FillMissingValues class
- Filter class
- FindIncrementalMatches class
- FindMatches class
- FlatMap class
- Join class
- Map class
- MapToCollection class
- mergeDynamicFrame
- Relationalize class
- RenameField class
- ResolveChoice class
- SelectFields class
- SelectFromCollection class

- Spigot class
- SplitFields class
- SplitRows class
- Unbox class
- UnnestFrame class

Setting up to use Python with AWS Glue

Use Python to develop your ETL scripts for Spark jobs. The supported Python versions for ETL jobs depend on the AWS Glue version of the job. For more information on AWS Glue versions, see the Glue version job property.

To set up your system for using Python with AWS Glue

Follow these steps to install Python and to be able to invoke the AWS Glue APIs.

1. If you don't already have Python installed, download and install it from the Python.org download page.
2. Install the AWS Command Line Interface (AWS CLI) as documented in the AWS CLI documentation.

   The AWS CLI is not directly necessary for using Python. However, installing and configuring it is a convenient way to set up AWS with your account credentials and verify that they work.
3. Install the AWS SDK for Python (Boto 3), as documented in the Boto3 Quickstart .

   Boto 3 resource APIs are not yet available for AWS Glue. Currently, only the Boto 3 client APIs can be used.

   For more information about Boto 3, see AWS SDK for Python (Boto3) Getting Started.

You can find Python code examples and utilities for AWS Glue in the AWS Glue samples repository on the GitHub website.

Calling AWS Glue APIs in Python

Note that Boto 3 resource APIs are not yet available for AWS Glue. Currently, only the Boto 3 client APIs can be used.

AWS Glue API names in Python

AWS Glue API names in Java and other programming languages are generally CamelCased. However, when called from Python, these generic names are changed to lowercase, with the parts of the name separated by underscore characters to make them more "Pythonic". In the *AWS Glue API* reference documentation, these Pythonic names are listed in parentheses after the generic CamelCased names.

However, although the AWS Glue API names themselves are transformed to lowercase, their parameter names remain capitalized. It is important to remember this, because parameters should be passed by name when calling AWS Glue APIs, as described in the following section.

Passing and accessing Python parameters in AWS Glue

In Python calls to AWS Glue APIs, it's best to pass parameters explicitly by name. For example:

```
job = glue.create_job(Name='sample', Role='Glue_DefaultRole',
                      Command={'Name': 'glueetl',
                               'ScriptLocation': 's3://my_script_bucket/scripts/
my_etl_script.py'})
```

It is helpful to understand that Python creates a dictionary of the name/value tuples that you specify as arguments to an ETL script in a Job structure or JobRun structure. Boto 3 then passes them to AWS Glue in JSON format by way of a REST API call. This means that you cannot rely on the order of the arguments when you access them in your script.

For example, suppose that you're starting a `JobRun` in a Python Lambda handler function, and you want to specify several parameters. Your code might look something like the following:

```
from datetime import datetime, timedelta

client = boto3.client('glue')

def lambda_handler(event, context):
  last_hour_date_time = datetime.now() - timedelta(hours = 1)
  day_partition_value = last_hour_date_time.strftime("%Y-%m-%d")
  hour_partition_value = last_hour_date_time.strftime("%-H")

  response = client.start_job_run(
               JobName = 'my_test_Job',
               Arguments = {
                 '--day_partition_key':   'partition_0',
                  '--hour_partition_key':  'partition_1',
                  '--day_partition_value':  day_partition_value,
                  '--hour_partition_value': hour_partition_value } )
```

To access these parameters reliably in your ETL script, specify them by name using AWS Glue's `getResolvedOptions` function and then access them from the resulting dictionary:

```
import sys
from awsglue.utils import getResolvedOptions

args = getResolvedOptions(sys.argv,
                          ['JOB_NAME',
                           'day_partition_key',
                           'hour_partition_key',
                           'day_partition_value',
                           'hour_partition_value'])
print "The day partition key is: ", args['day_partition_key']
print "and the day partition value is: ", args['day_partition_value']
```

If you want to pass an argument that is a nested JSON string, to preserve the parameter value as it gets passed to your AWS Glue ETL job, you must encode the parameter string before starting the job run, and then decode the parameter string before referencing it your job script. For example, consider the following argument string:

```
glue_client.start_job_run(JobName = "gluejobname", Arguments={
"--my_curly_braces_string": '{"a": {"b": {"c": [{"d": {"e": 42}}]}}}'
})
```

To pass this parameter correctly, you should encode the argument as a Base64 encoded string.

```
import base64
...
sample_string='{"a": {"b": {"c": [{"d": {"e": 42}}]}}}'
sample_string_bytes = sample_string.encode("ascii")

base64_bytes = base64.b64encode(sample_string_bytes)
base64_string = base64_bytes.decode("ascii")
...
glue_client.start_job_run(JobName = "gluejobname", Arguments={
"--my_curly_braces_string": base64_bytes})
...
sample_string_bytes = base64.b64decode(base64_bytes)
```

```
sample_string = sample_string_bytes.decode("ascii")
print(f"Decoded string: {sample_string}")
...
```

Example: Create and run a job

The following example shows how call the AWS Glue APIs using Python, to create and run an ETL job.

To create and run a job

1. Create an instance of the AWS Glue client:

   ```
   import boto3
   glue = boto3.client(service_name='glue', region_name='us-east-1',
                 endpoint_url='https://glue.us-east-1.amazonaws.com')
   ```

2. Create a job. You must use `glueetl` as the name for the ETL command, as shown in the following code:

   ```
   myJob = glue.create_job(Name='sample', Role='Glue_DefaultRole',
                             Command={'Name': 'glueetl',
                                      'ScriptLocation': 's3://my_script_bucket/
   scripts/my_etl_script.py'})
   ```

3. Start a new run of the job that you created in the previous step:

   ```
   myNewJobRun = glue.start_job_run(JobName=myJob['Name'])
   ```

4. Get the job status:

   ```
   status = glue.get_job_run(JobName=myJob['Name'], RunId=myNewJobRun['JobRunId'])
   ```

5. Print the current state of the job run:

   ```
   print(status['JobRun']['JobRunState'])
   ```

Using Python libraries with AWS Glue

You can install additional Python modules and libraries for use with AWS Glue ETL. For AWS Glue 2.0 and above, AWS Glue uses the Python Package Installer (pip3) to install additional modules used by AWS Glue ETL. AWS Glue provides multiple options to bring the additional Python modules to your AWS Glue job environment. You can use the “—additional-python-modules" parameter to bring in modules using Python wheel files, Requirements file (requirement.txt, AWS Glue 5.0 and above), or a list of comma-separated Python modules.

Topics

- Installing additional Python modules with pip in AWS Glue 2.0 or later
- Best Practices for Installing additional Python libraries in AWS Glue
- Including Python files with PySpark native features
- Programming scripts that use visual transforms
- Zipping libraries for inclusion
- Loading Python libraries in AWS Glue Studio notebooks
- Loading Python libraries in a development endpoint in AWS Glue 0.9/1.0
- Using Python libraries in a job or JobRun
- Proactively analyze Python dependencies
- Python modules already provided in AWS Glue

AWS Glue environment details

Glue Version Compatibility and Installation Methods

Glue version	Python version	Base image	glibc version	Supported Installation Methods
5.0	3.11	Amazon Linux 2023 (AL2023)	2.34	• the section called “(Recommended) Packaging Python environment into a single wheel file” • the section called “Installing additional Python libraries in AWS Glue 5.0 or above using requirements.txt” • the section called “Installing additional Python libraries using Wheel”

Glue version	Python version	Base image	glibc version	Supported Installation Methods
				- the section called "Installing additional Python modules with pip in AWS Glue 2.0 or later"
4.0	3.10	Amazon Linux 2 (AL2)	2.26	the section called " Best Practices for Installing additional Python libraries in AWS Glue "
3.0	3.7	Amazon Linux 2 (AL2)	2.26	the section called " Best Practices for Installing additional Python libraries in AWS Glue "
2.0	3.7	Amazon Linux AMI (AL1)	2.17	the section called " Best Practices for Installing additional Python libraries in AWS Glue "
1.0	3.6	Amazon Linux AMI (AL1)	2.17	the section called " Best Practices for Installing additional Python libraries in AWS Glue "
0.9	2.7	Amazon Linux AMI (AL1)	2.17	the section called " Best Practices for Installing additional Python libraries in AWS Glue "

Under the AWS shared responsibility model, you are responsible for the management of additional Python modules, libraries, and their dependencies that you use with your AWS Glue ETL jobs. This includes applying updates and security patches.

AWS Glue does not support compiling native code in the job environment. However, AWS Glue jobs run within an Amazon-managed Linux environment. You may be able to provide your native dependencies in a compiled form through a Python wheel file. Please refer to above table for AWS Glue version compatibility details.

If your Python dependencies transitively depend on native, compiled code, you may run against the following limitation: AWS Glue does not support compiling native code in the job environment. However, AWS Glue jobs run within an Amazon-managed Linux environment. You may be able to provide your native dependencies in a compiled form through a wheel distribution. Please refer to above table for AWS Glue version compatibility details.

> ⚠ **Important**
>
> Using incompatible dependencies can result in runtime issues, particularly for libraries with native extensions that must match the target environment's architecture and system libraries. Each AWS Glue version runs on a specific Python version with pre-installed libraries and system configurations.

Installing additional Python modules with pip in AWS Glue 2.0 or later

AWS Glue uses the Python Package Installer (pip3) to install additional modules to be used by AWS Glue ETL. You can use the `--additional-python-modules` parameter with a list of comma-separated Python modules to add a new module or change the version of an existing module. You can install custom distributions of a library by uploading the distribution to Amazon S3, then include the path to the Amazon S3 object in your list of modules.

You can pass additional options to pip3 with the `--python-modules-installer-option` parameter. For example, you could pass `"--upgrade"` to upgrade the packages specified by `"--additional-python-modules"`. For more examples, see Building Python modules from a wheel for Spark ETL workloads using AWS Glue 2.0 .

Installing additional Python libraries using Wheel

AWS Glue supports installing custom Python packages using wheel (.whl) files stored in Amazon S3. To include wheel files in your AWS Glue jobs, provide a comma-separated list of your wheel files stored in s3 to the `--additional-python-modules` job parameter. For example:

```
--additional-python-modules s3://amzn-s3-demo-bucket/path/to/package-1.0.0-py3-none-
any.whl,s3://your-bucket/path/to/another-package-2.1.0-cp311-cp311-linux_x86_64.whl
```

This approach also supports when you need custom distributions, or packages with native dependencies that are pre-compiled for the correct operating system. For more examples, see Building Python modules from a wheel for Spark ETL workloads using AWS Glue 2.0.

You specify the `--additional-python-modules` in the Job parameters field of the AWS Glue console or by altering the job arguments in the AWS SDK. For more information about setting job parameters, see Using job parameters in AWS Glue jobs.

Installing additional Python libraries in AWS Glue 5.0 or above using requirements.txt

In AWS Glue 5.0, you can provide the defacto-standard `requirements.txt` to manage Python library dependencies. To do that, provide following two job parameters:

- Key: `--python-modules-installer-option`

  Value: `-r`
- Key: `--additional-python-modules`

  Value: `s3://path_to_requirements.txt`

AWS Glue 5.0 nodes initially load python libraries specified in `requirements.txt`.

Here's a sample requirements.txt:

```
awswrangler==3.9.1
elasticsearch==8.15.1
PyAthena==3.9.0
PyMySQL==1.1.1
PyYAML==6.0.2
pyodbc==5.2.0
pyorc==0.9.0
redshift-connector==2.1.3
scipy==1.14.1
scikit-learn==1.5.2
SQLAlchemy==2.0.36
```

⚠ **Important**

Please avoid unpinned library versions in your requirements.txt to ensure you have a reliable and deterministic AWS Glue environment for your jobs.

When you use wheel for direct dependencies, you can bring in incompatible version of your transitive dependencies if they are not pinned correctly. As a best practice, all library versions should be pinned for consistency in AWS Glue jobs. AWS Glue recommends packaging your python environment into a wheel file to ensure consistency and reliability for your production workload.

Installing additional Python libraries directly configuring as comma separated list

To update or to add a new Python module AWS Glue allows passing `--additional-python-modules` parameter with a list of comma-separated Python modules as values. For example to update/ add scikit-learn module use the following key/value: `"--additional-python-modules", "scikit-learn==0.21.3"`. You have two options to directly configure the python modules.

- **Pinned Python Module (recommended)**

  ```
  "--additional-python-modules", "scikit-learn==0.21.3,ephem==4.1.6"
  ```
- **Unpinned Python Module: (Not recommended for Production Workloads)**

  ```
  "--additional-python-modules", "scikit-learn>==0.20.0,ephem>=4.0.0"
  ```

  OR

  ```
  "--additional-python-modules", "scikit-learn,ephem"
  ```

> ⚠ **Important**
>
> When configuring the python modules directly in `--additional-python-modules`, AWS Glue recommends to use pinned library versions to ensure consistency in AWS Glue job environment. Using unpinned library versions, pulls the latest version of the python modules, however this can introduce breaking changes or bring in incompatible python module leading to job failure due to python installation failure in AWS Glue job environment. We recommend customers to not use unpinned library versions for production workload. As a best practice, AWS Glue recommends packaging your python environment into a wheel file to ensure consistency and reliability for your production workload.

Best Practices for Installing additional Python libraries in AWS Glue

(Recommended) Packaging Python environment into a single wheel file

For a safe and consistent environment, AWS Glue recommends that you snapshot and package your python environment into a wheel file. The benefit of this is that your python environment for reference python modules and its transitive dependencies will be locked. This ensure that your

AWS Glue job is not affected when upstream repository like PyPI or dependencies introduces an incompatible updates.

This file can then be used in your AWS Glue job using the `--additional-python-modules` flag.

Important

You must run the following script in a similar environment as the AWS Glue version you are running. Please refer to the glue environment details table and ensure you are using the same base OS image and python version.

To package your libraries into an uber wheel file, you can use the following script:

```
#!/bin/bash
set -e
REQUIREMENTS_FILE="requirements.txt"
FINAL_WHEEL_OUTPUT_DIRECTORY="."
PACKAGE_NAME=$(basename "$(pwd)")
PACKAGE_VERSION="0.1.0"
# Help message
show_help() {
    echo "Usage: $0 [options]"
    echo ""
    echo "Options:"
    echo "  -r, --requirements FILE   Path to requirements.txt file (default:
 requirements.txt)"
    echo "  -o, --wheel-output DIR    Output directory for final wheel (default:
 current directory)"
    echo "  -n, --name NAME           Package name (default: current directory name)"
    echo "  -v, --version VERSION     Package version (default: 0.1.0)"
    echo "  -h, --help                Show this help message"
    echo "  -g, --glue-version        Glue version (required)"
    echo ""
    echo "Example:"
    echo "  $0 -r custom-requirements.txt -o dist -n my_package -v 1.2.3 -g 4.0"
}
# Parse command line arguments
while [[ $# -gt 0 ]]; do
    key="$1"
    case $key in
    -r | --requirements)
```

```
            REQUIREMENTS_FILE="$2"
            shift 2
            ;;
        -o | --wheel-output)
            FINAL_WHEEL_OUTPUT_DIRECTORY="$2"
            shift 2
            ;;
        -n | --name)
            PACKAGE_NAME="$2"
            shift 2
            ;;
        -v | --version)
            PACKAGE_VERSION="$2"
            shift 2
            ;;
        -g | --glue-version)
            GLUE_VERSION="$2"
            shift 2
            ;;
        -h | --help)
            show_help
            exit 0
            ;;
        *)
            echo "Unknown option: $1"
            show_help
            exit 1
            ;;
    esac
done
# If package name has dashes, convert to underscores and notify user. We need to check
 this since we cant import a package with dashes.
if [[ "$PACKAGE_NAME" =~ "-" ]]; then
    echo "Warning: Package name '$PACKAGE_NAME' contains dashes. Converting to
 underscores."
    PACKAGE_NAME=$(echo "$PACKAGE_NAME" | tr '-' '_')
fi
UBER_WHEEL_NAME="${PACKAGE_NAME}-${PACKAGE_VERSION}-py3-none-any.whl"
# Check if glue version is provided
if [ -z "$GLUE_VERSION" ]; then
    echo "Error: Glue version is required."
    exit 1
fi
# Validate version format (basic check)
```

```
if [[ ! "$PACKAGE_VERSION" =~ ^[0-9]+\.[0-9]+\.[0-9]+$ ]] && [[ ! "$PACKAGE_VERSION" =~
 ^[0-9]+\.[0-9]+$ ]]; then
    echo "Warning: Version '$PACKAGE_VERSION' doesn't follow semantic versioning (x.y.z
 or x.y)"
fi
# Check if requirements file exists
if [ ! -f "$REQUIREMENTS_FILE" ]; then
    echo "Error: Requirements file '$REQUIREMENTS_FILE' not found."
    exit 1
fi
# Get relevant platform tags/python versions based on glue version
if [[ "$GLUE_VERSION" == "5.0" ]]; then
    PYTHON_VERSION="3.11"
    GLIBC_VERSION="2.34"
elif [[ "$GLUE_VERSION" == "4.0" ]]; then
    PYTHON_VERSION="3.10"
    GLIBC_VERSION="2.26"
elif [[ "$GLUE_VERSION" == "3.0" ]]; then
    PYTHON_VERSION="3.7"
    GLIBC_VERSION="2.26"
elif [[ "$GLUE_VERSION" == "2.0" ]]; then
    PYTHON_VERSION="3.7"
    GLIBC_VERSION="2.17"
elif [[ "$GLUE_VERSION" == "1.0" ]]; then
    PYTHON_VERSION="3.6"
    GLIBC_VERSION="2.17"
elif [[ "$GLUE_VERSION" == "0.9" ]]; then
    PYTHON_VERSION="2.7"
    GLIBC_VERSION="2.17"
else
    echo "Error: Unsupported glue version '$GLUE_VERSION'."
    exit 1
fi
echo "Using Glue version $GLUE_VERSION"
echo "Using Glue python version $PYTHON_VERSION"
echo "Using Glue glibc version $GLIBC_VERSION"
PIP_PLATFORM_FLAG=""
is_glibc_compatible() {
    # assumes glibc version in the form of major.minor (ex: 2.17)
    # glue glibc must be >= platform glibc
    local glue_glibc_version="$GLIBC_VERSION"
    local platform_glibc_version="$1"
    # 2.27 (platform) can run on 2.27 (glue)
    if [[ "$platform_glibc_version" == "$glue_glibc_version" ]]; then
```

```
        return 0
    fi
    local glue_glibc_major="${glue_glibc_version%%.*}"
    local glue_glibc_minor="${glue_glibc_version#*.}"
    local platform_glibc_major="${platform_glibc_version%%.*}"
    local platform_glibc_minor="${platform_glibc_version#*.}"
    # 3.27 (platform) cannot run on 2.27 (glue)
    if [[ "$platform_glibc_major" -gt "$glue_glibc_major" ]]; then
        return 1
    fi
    # 2.34 (platform) cannot run on 2.27 (glue)
    if [[ "$platform_glibc_major" -eq "$glue_glibc_major" ]] &&
 [[ "$platform_glibc_minor" -gt "$glue_glibc_minor" ]]; then
        return 1
    fi
    # 2.17 (platform) can run on 2.27 (glue)
    return 0
}
PIP_PLATFORM_FLAG=""
if is_glibc_compatible "2.17"; then
    PIP_PLATFORM_FLAG="${PIP_PLATFORM_FLAG} --platform manylinux2014_x86_64"
fi
if is_glibc_compatible "2.28"; then
    PIP_PLATFORM_FLAG="${PIP_PLATFORM_FLAG} --platform manylinux_2_28_x86_64"
fi
if is_glibc_compatible "2.34"; then
    PIP_PLATFORM_FLAG="${PIP_PLATFORM_FLAG} --platform manylinux_2_34_x86_64"
fi
if is_glibc_compatible "2.39"; then
    PIP_PLATFORM_FLAG="${PIP_PLATFORM_FLAG} --platform manylinux_2_39_x86_64"
fi
echo "Using pip platform flags: $PIP_PLATFORM_FLAG"
# Convert to absolute paths
REQUIREMENTS_FILE=$(realpath "$REQUIREMENTS_FILE")
FINAL_WHEEL_OUTPUT_DIRECTORY=$(realpath "$FINAL_WHEEL_OUTPUT_DIRECTORY")
TEMP_WORKING_DIR=$(mktemp -d)
VENV_DIR="${TEMP_WORKING_DIR}/.build_venv"
WHEEL_OUTPUT_DIRECTORY="${TEMP_WORKING_DIR}/wheelhouse"
# Cleanup function
cleanup() {
    echo "Cleaning up temporary files..."
    rm -rf "$TEMP_WORKING_DIR"
}
trap cleanup EXIT
```

```
echo "========================================="
echo "Building wheel for $PACKAGE_NAME with all dependencies from $REQUIREMENTS_FILE"
echo "========================================="
# Determine Python executable to use consistently
PYTHON_EXEC=$(which python3 2>/dev/null || which python 2>/dev/null)
if [ -z "$PYTHON_EXEC" ]; then
    echo "Error: No Python executable found"
    exit 1
fi
echo "Using Python: $PYTHON_EXEC"
echo ""
# Install build requirements
echo "Step 1/5: Installing build tools..."
echo "----------------------------------------"
"$PYTHON_EXEC" -m pip install --upgrade pip build wheel setuptools
echo "# Build tools installed successfully"
echo ""
# Create a virtual environment for building
echo "Step 2/5: Creating build environment..."
echo "----------------------------------------"
"$PYTHON_EXEC" -m venv "$VENV_DIR"
# Check if virtual environment was created successfully
if [ ! -f "$VENV_DIR/bin/activate" ]; then
    echo "Error: Failed to create virtual environment"
    exit 1
fi
source "$VENV_DIR/bin/activate"
# Install pip-tools for dependency resolution
"$VENV_DIR/bin/pip" install pip-tools
echo "# Build environment created successfully"
echo ""
# Compile requirements to get all transitive dependencies
GLUE_PIP_ARGS="$PIP_PLATFORM_FLAG --python-version $PYTHON_VERSION --only-binary=:all:"
echo "Step 3/5: Resolving all dependencies..."
echo "----------------------------------------"
if ! "$VENV_DIR/bin/pip-compile" --pip-args "$GLUE_PIP_ARGS" --no-emit-index-url --
output-file "$TEMP_WORKING_DIR/.compiled_requirements.txt" "$REQUIREMENTS_FILE"; then
    echo "Error: Failed to resolve dependencies. Check for conflicts in
 $REQUIREMENTS_FILE"
    exit 1
fi
echo "# Dependencies resolved successfully"
echo ""
# Download all wheels for dependencies
```

```
echo "Step 4/5: Downloading all dependency wheels..."
echo "----------------------------------------"
"$VENV_DIR/bin/pip" download -r "$TEMP_WORKING_DIR/.compiled_requirements.txt" -d
 "$WHEEL_OUTPUT_DIRECTORY" $GLUE_PIP_ARGS
# Check if any wheels were downloaded
if [ ! "$(ls -A "$WHEEL_OUTPUT_DIRECTORY")" ]; then
    echo "Error: No wheels were downloaded. Check your requirements file."
    exit 1
fi
# Count downloaded wheels (using find instead of ls for better handling)
WHEEL_COUNT=$(find "$WHEEL_OUTPUT_DIRECTORY" -name "*.whl" -type f | wc -l | tr -d ' ')
echo "# Downloaded $WHEEL_COUNT dependency wheels successfully"
echo ""
# Create a single uber wheel with all dependencies
echo "Step 5/5: Creating uber wheel with all dependencies included..."
echo "----------------------------------------"
# Create a temporary directory for the uber wheel
UBER_WHEEL_DIR="$TEMP_WORKING_DIR/uber"
mkdir -p "$UBER_WHEEL_DIR"
# Create the setup.py file with custom install command
cat >"$UBER_WHEEL_DIR/setup.py" <<EOF
from setuptools import setup, find_packages
import setuptools.command.install
import os
import glob
import subprocess
import sys
setup(
    name='${PACKAGE_NAME}',
    version='${PACKAGE_VERSION}',
    description='Bundle containing dependencies for ${PACKAGE_NAME}',
    author='Package Builder',
    author_email='builder@example.com',
    packages=['${PACKAGE_NAME}'],  # Include the package directory to hold wheels
    include_package_data=True,
    package_data={
        '${PACKAGE_NAME}': ['wheels/*.whl'],  # Include wheels in the package directory
    }
)
EOF
# Create a MANIFEST.in file to include all wheels
cat >"$UBER_WHEEL_DIR/MANIFEST.in" <<EOF
recursive-include ${PACKAGE_NAME}/wheels *.whl
EOF
```

```
# Create an __init__.py file that imports all the bundled wheel files (no auto-install
 logic)
mkdir -p "$UBER_WHEEL_DIR/${PACKAGE_NAME}"
cat >"$UBER_WHEEL_DIR/${PACKAGE_NAME}/__init__.py" <<EOF
"""
${PACKAGE_NAME} - dependencies can be installed at runtime using the $(load_wheels)
 function
"""
from pathlib import Path
import logging
import subprocess
import sys
__version__ = "${PACKAGE_VERSION}"

def load_wheels(log_level=logging.INFO):
    logger = logging.getLogger(__name__)
    handler = logging.StreamHandler(sys.stdout)
    formatter = logging.Formatter("[Glue Python Wheel Installer] %(asctime)s - %(name)s
 - %(levelname)s - %(message)s")
    handler.setFormatter(formatter)
    logger.addHandler(handler)
    logger.setLevel(log_level)
    logger.info("Starting wheel installation process")
    package_dir = Path(__file__).parent.absolute()
    wheels_dir = package_dir / "wheels"
    logger.debug(f"Package directory: {package_dir}")
    logger.debug(f"Looking for wheels in: {wheels_dir}")
    if not wheels_dir.exists():
        logger.error(f"Wheels directory not found: {wheels_dir}")
        return False
    wheel_files = list(wheels_dir.glob("*.whl"))
    if not wheel_files:
        logger.warning(f"No wheels found in: {wheels_dir}")
        return False
    logger.info(f"Found {len(wheel_files)} wheels")
    wheel_file_paths = [str(wheel_file) for wheel_file in wheel_files]
    logger.info(f"Installing {wheel_file_paths}...")
    try:
        result = subprocess.run(
            [sys.executable, "-m", "pip", "install", *wheel_file_paths], check=True,
 capture_output=True, text=True
        )
        logger.info(f"# Successfully installed wheel files")
        logger.debug(f"pip output: {result.stdout}")
    except subprocess.CalledProcessError as e:
        error_msg = f"Failed to install wheel files"
        logger.error(f"# {error_msg}: {e}")
        if e.stderr:
            logger.error(f"Error details: {e.stderr}")
        return False
    logger.info("All wheels installed successfully")
    return True
EOF
cat >"$UBER_WHEEL_DIR/${PACKAGE_NAME}/auto.py" <<EOF
"""
${PACKAGE_NAME} - utility module that allows users to automatically install modules by
 adding $(import ${PACKAGE_NAME}.auto) to the top of their script
"""
from ${PACKAGE_NAME} import load_wheels
load_wheels()
EOF
# Copy all wheels to the uber wheel directory
mkdir -p "$UBER_WHEEL_DIR/${PACKAGE_NAME}/wheels"
cp "$WHEEL_OUTPUT_DIRECTORY"/*.whl "$UBER_WHEEL_DIR/${PACKAGE_NAME}/wheels/"
# Build the uber wheel
echo "Building uber wheel package..."
# Install build tools in the current environment
"$VENV_DIR/bin/pip" install build
if ! (cd "$UBER_WHEEL_DIR" && "$VENV_DIR/bin/python" -m build --skip-dependency-check
 --wheel --outdir .); then
    echo "Error: Failed to build uber wheel"
    exit 1
fi
# Ensure output directory exists
mkdir -p "$FINAL_WHEEL_OUTPUT_DIRECTORY"
# Copy the uber wheel to the output directory
FINAL_WHEEL_OUTPUT_PATH="$FINAL_WHEEL_OUTPUT_DIRECTORY/$UBER_WHEEL_NAME"
# Find the generated wheel (should be only one in the root directory)
GENERATED_WHEEL=$(find "$UBER_WHEEL_DIR" -maxdepth 1 -name "*.whl" -type f | head -1)
if [ -z "$GENERATED_WHEEL" ]; then
    echo "Error: No uber wheel was generated"
    exit 1
fi
cp "$GENERATED_WHEEL" "$FINAL_WHEEL_OUTPUT_PATH"
# Get final wheel size for user feedback
WHEEL_SIZE=$(du -h "$FINAL_WHEEL_OUTPUT_PATH" | cut -f1)
echo "# Uber wheel created successfully!"
echo ""
echo "=========================================="
echo "BUILD COMPLETED SUCCESSFULLY!"
echo "=========================================="
echo "Final wheel: $FINAL_WHEEL_OUTPUT_PATH"
echo "Wheel size: $WHEEL_SIZE"
echo "Dependencies included: $WHEEL_COUNT packages"
echo ""
echo "To install the bundle, run:"
echo "  pip install $FINAL_WHEEL_OUTPUT_PATH"
echo ""
echo "After installation, you can verify that the bundle works by running:"
echo "  python -c \"import ${PACKAGE_NAME}; ${PACKAGE_NAME}.load_wheels()\""
echo "  or "
echo "  python -c \"import ${PACKAGE_NAME}.auto\""
echo "=========================================="
```

You can run this tool with the following command:

```
./wheel_packager.sh -r <path to requirements.txt> -g <glue version> -o <wheel output
 directory> -n <package name> -v <wheel version>
```

To include this wheel file in your AWS Glue jobs, include the Amazon S3 location of the file in the `--additional-python-modules` job parameter. For Example:

```
--additional-python-modules s3://your-bucket/path/to/package_with_dependencies-1.0.0-
py3-none-any.whl
```

Then at the top of your python script, you can install your bundled dependencies at runtime.

```
# Option 1: automatic installation via import
import package_with_dependencies.auto

# Option 2: manual installation
from package_with_dependencies import load_wheels
load_wheels()
```

Including Python files with PySpark native features

AWS Glue uses PySpark to include Python files in AWS Glue ETL jobs. You will want to use `--additional-python-modules` to manage your dependencies when available. You can use the `--`

`extra-py-files` job parameter to include Python files. Dependencies must be hosted in Amazon S3 and the argument value should be a comma delimited list of Amazon S3 paths with no spaces. This functionality behaves like the Python dependency management you would use with Spark. For more information on Python dependency management in Spark, see Using PySpark Native Features page in Apache Spark documentation. `--extra-py-files` is useful in cases where your additional code is not packaged, or when you are migrating a Spark program with an existing toolchain for managing dependencies. For your dependency tooling to be maintainable, you will have to bundle your dependencies before submitting.

Programming scripts that use visual transforms

When you create a AWS Glue job using the AWS Glue Studio visual interface, you can transform your data with managed data transform nodes and custom visual transforms. For more information about managed data transform nodes, see the section called “Transform data with AWS Glue managed transforms”. For more information about custom visual transforms, see the section called “ Transform data with custom visual transforms ”. Scripts using visual transforms can only be generated when your job **Language** is set to use Python.

When generating a AWS Glue job using visual transforms, AWS Glue Studio will include these transforms in the runtime environment using the `--extra-py-files` parameter in the job configuration. For more information about job parameters, see the section called “Job parameters”. When making changes to a generated script or runtime environment, you will need to preserve this job configuration for your script to run successfully.

Zipping libraries for inclusion

Unless a library is contained in a single `.py` file, it should be packaged in a `.zip` archive. The package directory should be at the root of the archive, and must contain an `__init__.py` file for the package. Python will then be able to import the package in the normal way.

If your library only consists of a single Python module in one `.py` file, you do not need to place it in a `.zip` file.

Loading Python libraries in AWS Glue Studio notebooks

To specify Python libraries in AWS Glue Studio notebooks, see Installing additional Python modules.

Loading Python libraries in a development endpoint in AWS Glue 0.9/1.0

If you are using different library sets for different ETL scripts, you can either set up a separate development endpoint for each set, or you can overwrite the library `.zip` file(s) that your development endpoint loads every time you switch scripts.

You can use the console to specify one or more library .zip files for a development endpoint when you create it. After assigning a name and an IAM role, choose **Script Libraries and job parameters (optional)** and enter the full Amazon S3 path to your library `.zip` file in the **Python library path** box. For example:

```
s3://bucket/prefix/site-packages.zip
```

If you want, you can specify multiple full paths to files, separating them with commas but no spaces, like this:

```
s3://bucket/prefix/lib_A.zip,s3://bucket_B/prefix/lib_X.zip
```

If you update these `.zip` files later, you can use the console to re-import them into your development endpoint. Navigate to the developer endpoint in question, check the box beside it, and choose **Update ETL libraries** from the **Action** menu.

In a similar way, you can specify library files using the AWS Glue APIs. When you create a development endpoint by calling CreateDevEndpoint action (Python: create_dev_endpoint), you can specify one or more full paths to libraries in the `ExtraPythonLibsS3Path` parameter, in a call that looks this:

```

dep = glue.create_dev_endpoint(
             EndpointName="testDevEndpoint",
             RoleArn="arn:aws:iam::123456789012",
             SecurityGroupIds="sg-7f5ad1ff",
             SubnetId="subnet-c12fdba4",
             PublicKey="ssh-rsa  AAAAB3NzaC1yc2EAAAADAQABAAABAQCtp04H/y...",
             NumberOfNodes=3,
             ExtraPythonLibsS3Path="s3://bucket/prefix/lib_A.zip,s3://bucket_B/prefix/
lib_X.zip")
```

When you update a development endpoint, you can also update the libraries it loads using a DevEndpointCustomLibraries object and setting the `UpdateEtlLibraries` parameter to `True` when calling UpdateDevEndpoint (update_dev_endpoint).

Using Python libraries in a job or JobRun

When you are creating a new Job on the console, you can specify one or more library .zip files by choosing **Script Libraries and job parameters (optional)** and entering the full Amazon S3 library path(s) in the same way you would when creating a development endpoint:

```
s3://bucket/prefix/lib_A.zip,s3://bucket_B/prefix/lib_X.zip
```

If you are calling CreateJob (create_job), you can specify one or more full paths to default libraries using the `--extra-py-files` default parameter, like this:

```
job = glue.create_job(Name='sampleJob',
                      Role='Glue_DefaultRole',
                      Command={'Name': 'glueetl',
                               'ScriptLocation': 's3://my_script_bucket/scripts/
my_etl_script.py'},
                      DefaultArguments={'--extra-py-files': 's3://bucket/prefix/
lib_A.zip,s3://bucket_B/prefix/lib_X.zip'})
```

Then when you are starting a JobRun, you can override the default library setting with a different one:

```
runId = glue.start_job_run(JobName='sampleJob',
                           Arguments={'--extra-py-files': 's3://bucket/prefix/
lib_B.zip'})
```

Proactively analyze Python dependencies

To proactively identify potential dependency issues before deploying to AWS Glue, you can use the dependency analysis tool to validate your Python packages against your target AWS Glue environment.

AWS provides an open-source Python dependency analyzer tool specifically designed for AWS Glue environments. This tool is available in the AWS Glue samples repository and can be used locally to validate your dependencies before deployment.

This analysis helps ensure your dependencies follow the recommended practice of pinning all library versions for consistent production deployments. For more details, please see the tool's README .

Using the AWS Glue Dependency Analyzer

The AWS Glue Python Dependency Analyzer helps identify unpinned dependencies and version conflicts by simulating pip installation with platform-specific constraints that match your target AWS Glue environment.

```
# Analyze a single Glue job
python glue_dependency_analyzer.py -j my-glue-job

# Analyze multiple jobs with specific AWS configuration
python glue_dependency_analyzer.py -j job1 -j job2 --aws-profile production --aws-
region us-west-2
```

The tool will flag:

- Unpinned dependencies that could install different versions across job runs
- Version conflicts between packages
- Dependencies not available for your target AWS Glue environment

Analyze and fix job failures due to Python dependencies with Amazon Q Developer

Amazon Q Developer is a generative artificial intelligence (AI) powered conversational assistant that can help you understand, build, extend, and operate AWS applications. You can download it by following the instructions in the Getting started guide for Amazon Q.

Amazon Q Developer can be used to analyze and fix job failures due to python dependency. We suggest using the following prompt by replacing the job <Job-Name> placeholder with the name of your glue job.

```
I have an AWS Glue job named <Job-Name> that has failed due to Python module
 installation conflicts. Please assist in diagnosing and resolving this issue using the
 following systematic approach. Proceed once sufficient information is available.

Objective: Implement a fix that addresses the root cause module while minimizing
 disruption to the existing working environment.
```

```
Step 1: Root Cause Analysis
• Retrieve the most recent failed job run ID for the specified Glue job
• Extract error logs from CloudWatch Logs using the job run ID as a log stream prefix
• Analyze the logs to identify:
  • The recently added or modified Python module that triggered the dependency conflict
  • The specific dependency chain causing the installation failure
  • Version compatibility conflicts between required and existing modules

Step 2: Baseline Configuration Identification
• Locate the last successful job run ID prior to the dependency failure
• Document the Python module versions that were functioning correctly in that baseline
 run
• Establish the compatible version constraints for conflicting dependencies

Step 3: Targeted Resolution Implementation
• Apply pinning by updating the job's additional_python_modules parameter
• Pin only the root cause module and its directly conflicting dependencies to
 compatible versions, and do not remove python modules unless necessary
• Preserve flexibility for non-conflicting modules by avoiding unnecessary version
 constraints
• Deploy the configuration changes with minimal changes to the existing configuration
 and execute a validation test run. Do not change the Glue versions.

Implementation Example:
Scenario: Recently added pandas==2.0.0 to additional_python_modules
Error: numpy version conflict (pandas 2.0.0 requires numpy>=1.21, but existing job code
 requires numpy<1.20)
Resolution: Update additional_python_modules to "pandas==1.5.3,numpy==1.19.5"
Rationale: Use pandas 1.5.3 (compatible with numpy 1.19.5) and pin numpy to last known
 working version

Expected Outcome: Restore job functionality with minimal configuration changes while
 maintaining system stability.
```

The prompt instructs Q to:

1. Fetch the latest failed job run ID
2. Find associated logs and details
3. Find successful job runs to detect any changed Python packages
4. Make any configuration fixes and trigger another test run

Python modules already provided in AWS Glue

To change the version of these provided modules, provide new versions with the `--additional-python-modules` job parameter.

AWS Glue version 5.0

AWS Glue version 5.0 includes the following Python modules out of the box:

- aiobotocore==2.13.1
- aiohappyeyeballs==2.3.5
- aiohttp==3.10.1
- aioitertools==0.11.0
- aiosignal==1.3.1
- appdirs==1.4.4
- attrs==24.2.0
- boto3==1.34.131
- botocore==1.34.131
- certifi==2024.7.4
- charset-normalizer==3.3.2
- contourpy==1.2.1
- cycler==0.12.1
- fonttools==4.53.1
- frozenlist==1.4.1
- fsspec==2024.6.1
- idna==2.10
- jmespath==0.10.0
- kaleido==0.2.1
- kiwisolver==1.4.5
- matplotlib==3.9.0
- multidict==6.0.5
- numpy==1.26.4

- packaging==24.1
- pandas==2.2.2
- pillow==10.4.0
- pip==23.0.1
- plotly==5.23.0
- pyarrow==17.0.0
- pyparsing==3.1.2
- python-dateutil==2.9.0.post0
- pytz==2024.1
- requests==2.32.2
- s3fs==2024.6.1
- s3transfer==0.10.2
- seaborn==0.13.2
- setuptools==59.6.0
- six==1.16.0
- tenacity==9.0.0
- tzdata==2024.1
- urllib3==1.25.10
- virtualenv==20.4.0
- wrapt==1.16.0
- yarl==1.9.4

AWS Glue version 4.0

AWS Glue version 4.0 includes the following Python modules out of the box:

- aiobotocore==2.4.1
- aiohttp==3.8.3
- aioitertools==0.11.0
- aiosignal==1.3.1
- async-timeout==4.0.2

- asynctest==0.13.0
- attrs==22.2.0
- avro-python3==1.10.2
- boto3==1.24.70
- botocore==1.27.59
- certifi==2021.5.30
- chardet==3.0.4
- charset-normalizer==2.1.1
- click==8.1.3
- cycler==0.10.0
- Cython==0.29.32
- fsspec==2021.8.1
- idna==2.10
- importlib-metadata==5.0.0
- jmespath==0.10.0
- joblib==1.0.1
- kaleido==0.2.1
- kiwisolver==1.4.4
- matplotlib==3.4.3
- mpmath==1.2.1
- multidict==6.0.4
- nltk==3.7
- numpy==1.23.5
- packaging==23.0
- pandas==1.5.1
- patsy==0.5.1
- Pillow==9.4.0
- pip==23.0.1
- plotly==5.16.0

- pmdarima==2.0.1
- ptvsd==4.3.2
- pyarrow==10.0.0
- pydevd==2.5.0
- pyhocon==0.3.58
- PyMySQL==1.0.2
- pyparsing==2.4.7
- python-dateutil==2.8.2
- pytz==2021.1
- PyYAML==6.0.1
- regex==2022.10.31
- requests==2.23.0
- s3fs==2022.11.0
- s3transfer==0.6.0
- scikit-learn==1.1.3
- scipy==1.9.3
- setuptools==49.1.3
- six==1.16.0
- statsmodels==0.13.5
- subprocess32==3.5.4
- sympy==1.8
- tbats==1.1.0
- threadpoolctl==3.1.0
- tqdm==4.64.1
- typing_extensions==4.4.0
- urllib3==1.25.11
- wheel==0.37.0
- wrapt==1.14.1
- yarl==1.8.2

- zipp==3.10.0

AWS Glue version 3.0

AWS Glue version 3.0 includes the following Python modules out of the box:,

- aiobotocore==1.4.2
- aiohttp==3.8.3
- aioitertools==0.11.0
- aiosignal==1.3.1
- async-timeout==4.0.2
- asynctest==0.13.0
- attrs==22.2.0
- avro-python3==1.10.2
- boto3==1.18.50
- botocore==1.21.50
- certifi==2021.5.30
- chardet==3.0.4
- charset-normalizer==2.1.1
- click==8.1.3
- cycler==0.10.0
- Cython==0.29.4
- docutils==0.17.1
- enum34==1.1.10
- frozenlist==1.3.3
- fsspec==2021.8.1
- idna==2.10
- importlib-metadata==6.0.0
- jmespath==0.10.0
- joblib==1.0.1
- kiwisolver==1.3.2

- matplotlib==3.4.3
- mpmath==1.2.1
- multidict==6.0.4
- nltk==3.6.3
- numpy==1.19.5
- packaging==23.0
- pandas==1.3.2
- patsy==0.5.1
- Pillow==9.4.0
- pip==23.0
- pmdarima==1.8.2
- ptvsd==4.3.2
- pyarrow==5.0.0
- pydevd==2.5.0
- pyhocon==0.3.58
- PyMySQL==1.0.2
- pyparsing==2.4.7
- python-dateutil==2.8.2
- pytz==2021.1
- PyYAML==5.4.1
- regex==2022.10.31
- requests==2.23.0
- s3fs==2021.8.1
- s3transfer==0.5.0
- scikit-learn==0.24.2
- scipy==1.7.1
- six==1.16.0
- Spark==1.0
- statsmodels==0.12.2
- subprocess32==3.5.4

- sympy==1.8
- tbats==1.1.0
- threadpoolctl==3.1.0
- tqdm==4.64.1
- typing_extensions==4.4.0
- urllib3==1.25.11
- wheel==0.37.0
- wrapt==1.14.1
- yarl==1.8.2
- zipp==3.12.0

AWS Glue version 2.0

AWS Glue version 2.0 includes the following Python modules out of the box:

- avro-python3==1.10.0
- awscli==1.27.60
- boto3==1.12.4
- botocore==1.15.4
- certifi==2019.11.28
- chardet==3.0.4
- click==8.1.3
- colorama==0.4.4
- cycler==0.10.0
- Cython==0.29.15
- docutils==0.15.2
- enum34==1.1.9
- fsspec==0.6.2
- idna==2.9
- importlib-metadata==6.0.0
- jmespath==0.9.4

- joblib==0.14.1
- kiwisolver==1.1.0
- matplotlib==3.1.3
- mpmath==1.1.0
- nltk==3.5
- numpy==1.18.1
- pandas==1.0.1
- patsy==0.5.1
- pmdarima==1.5.3
- ptvsd==4.3.2
- pyarrow==0.16.0
- pyasn1==0.4.8
- pydevd==1.9.0
- pyhocon==0.3.54
- PyMySQL==0.9.3
- pyparsing==2.4.6
- python-dateutil==2.8.1
- pytz==2019.3
- PyYAML==5.3.1
- regex==2022.10.31
- requests==2.23.0
- rsa==4.7.2
- s3fs==0.4.0
- s3transfer==0.3.3
- scikit-learn==0.22.1
- scipy==1.4.1
- setuptools==45.2.0
- six==1.14.0
- Spark==1.0
- statsmodels==0.11.1

- subprocess32==3.5.4
- sympy==1.5.1
- tbats==1.0.9
- tqdm==4.64.1
- typing-extensions==4.4.0
- urllib3==1.25.8
- wheel==0.35.1
- zipp==3.12.0

AWS Glue Python code samples

- Code example: Joining and relationalizing data
- Code example: Data preparation using ResolveChoice, Lambda, and ApplyMapping

Code example: Joining and relationalizing data

This example uses a dataset that was downloaded from http://everypolitician.org/ to the `sample-dataset` bucket in Amazon Simple Storage Service (Amazon S3): `s3://awsglue-datasets/examples/us-legislators/all`. The dataset contains data in JSON format about United States legislators and the seats that they have held in the US House of Representatives and Senate, and has been modified slightly and made available in a public Amazon S3 bucket for purposes of this tutorial.

You can find the source code for this example in the `join_and_relationalize.py` file in the AWS Glue samples repository on the GitHub website.

Using this data, this tutorial shows you how to do the following:

- Use an AWS Glue crawler to classify objects that are stored in a public Amazon S3 bucket and save their schemas into the AWS Glue Data Catalog.
- Examine the table metadata and schemas that result from the crawl.
- Write a Python extract, transfer, and load (ETL) script that uses the metadata in the Data Catalog to do the following:
  - Join the data in the different source files together into a single data table (that is, denormalize the data).

- Filter the joined table into separate tables by type of legislator.
- Write out the resulting data to separate Apache Parquet files for later analysis.

The preferred way to debug Python or PySpark scripts while running on AWS is to use Notebooks on AWS Glue Studio.

Step 1: Crawl the data in the Amazon S3 bucket

1. Sign in to the AWS Management Console, and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. Following the steps in Configuring a crawler, create a new crawler that can crawl the `s3://awsglue-datasets/examples/us-legislators/all` dataset into a database named `legislators` in the AWS Glue Data Catalog. The example data is already in this public Amazon S3 bucket.
3. Run the new crawler, and then check the `legislators` database.

   The crawler creates the following metadata tables:

   - `persons_json`
   - `memberships_json`
   - `organizations_json`
   - `events_json`
   - `areas_json`
   - `countries_r_json`

   This is a semi-normalized collection of tables containing legislators and their histories.

Step 2: Add boilerplate script to the development endpoint notebook

Paste the following boilerplate script into the development endpoint notebook to import the AWS Glue libraries that you need, and set up a single `GlueContext`:

```
import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job

glueContext = GlueContext(SparkContext.getOrCreate())
```

Step 3: Examine the schemas from the data in the Data Catalog

Next, you can easily create examine a DynamicFrame from the AWS Glue Data Catalog, and examine the schemas of the data. For example, to see the schema of the `persons_json` table, add the following in your notebook:

```

persons = glueContext.create_dynamic_frame.from_catalog(
             database="legislators",
             table_name="persons_json")
print "Count: ", persons.count()
persons.printSchema()
```

Here's the output from the print calls:

```

Count:  1961
root
|-- family_name: string
|-- name: string
|-- links: array
|    |-- element: struct
|    |    |-- note: string
|    |    |-- url: string
|-- gender: string
|-- image: string
|-- identifiers: array
|    |-- element: struct
|    |    |-- scheme: string
|    |    |-- identifier: string
|-- other_names: array
|    |-- element: struct
|    |    |-- note: string
|    |    |-- name: string
|    |    |-- lang: string
|-- sort_name: string
|-- images: array
|    |-- element: struct
|    |    |-- url: string
|-- given_name: string
|-- birth_date: string
|-- id: string
|-- contact_details: array
|    |-- element: struct
|    |    |-- type: string
|    |    |-- value: string
|-- death_date: string
```

Each person in the table is a member of some US congressional body.

To view the schema of the `memberships_json` table, type the following:

```
memberships = glueContext.create_dynamic_frame.from_catalog(
                 database="legislators",
                 table_name="memberships_json")
print "Count: ", memberships.count()
memberships.printSchema()
```

The output is as follows:

```
Count:  10439
root
|-- area_id: string
|-- on_behalf_of_id: string
|-- organization_id: string
|-- role: string
|-- person_id: string
|-- legislative_period_id: string
|-- start_date: string
|-- end_date: string
```

The `organizations` are parties and the two chambers of Congress, the Senate and House of Representatives. To view the schema of the `organizations_json` table, type the following:

```
orgs = glueContext.create_dynamic_frame.from_catalog(
            database="legislators",
            table_name="organizations_json")
print "Count: ", orgs.count()
orgs.printSchema()
```

The output is as follows:

```

Count:  13
root
|-- classification: string
|-- links: array
|    |-- element: struct
|    |    |-- note: string
|    |    |-- url: string
|-- image: string
|-- identifiers: array
|    |-- element: struct
|    |    |-- scheme: string
|    |    |-- identifier: string
|-- other_names: array
|    |-- element: struct
|    |    |-- lang: string
|    |    |-- note: string
|    |    |-- name: string
|-- id: string
|-- name: string
|-- seats: int
|-- type: string
```

Step 4: Filter the data

Next, keep only the fields that you want, and rename `id` to `org_id`. The dataset is small enough that you can view the whole thing.

The `toDF()` converts a `DynamicFrame` to an Apache Spark `DataFrame`, so you can apply the transforms that already exist in Apache Spark SQL:

```

orgs = orgs.drop_fields(['other_names',
                        'identifiers']).rename_field(
                            'id', 'org_id').rename_field(
                               'name', 'org_name')
```

```
orgs.toDF().show()
```

The following shows the output:

```
+--------------+--------------------+--------------------+--------------------+-----
+-----------+--------------------+
|classification|              org_id|            org_name|               links|seats|
     type|               image|
+--------------+--------------------+--------------------+--------------------+-----
+-----------+--------------------+
|         party|            party/al|                  AL|                null| null|
     null|                null|
|         party|      party/democrat|            Democrat|[[website,http://...| null|
     null|https://upload.wi...|
|         party|party/democrat-li...|    Democrat-Liberal|[[website,http://...| null|
     null|                null|
|   legislature|d56acebe-8fdc-47b...|House of Represen...|                null|  435|
lower house|                null|
|         party|   party/independent|         Independent|                null| null|
     null|                null|
|         party|party/new_progres...|     New Progressive|[[website,http://...| null|
     null|https://upload.wi...|
|         party|party/popular_dem...|    Popular Democrat|[[website,http://...| null|
     null|                null|
|         party|    party/republican|          Republican|[[website,http://...| null|
     null|https://upload.wi...|
|         party|party/republican-...|Republican-Conser...|[[website,http://...| null|
     null|                null|
|         party|      party/democrat|            Democrat|[[website,http://...| null|
     null|https://upload.wi...|
|         party|   party/independent|         Independent|                null| null|
     null|                null|
|         party|    party/republican|          Republican|[[website,http://...| null|
     null|https://upload.wi...|
|   legislature|8fa6c3d2-71dc-478...|              Senate|                null|  100|
upper house|                null|
+--------------+--------------------+--------------------+--------------------+-----
+-----------+--------------------+
```

Type the following to view the `organizations` that appear in `memberships`:

```
memberships.select_fields(['organization_id']).toDF().distinct().show()
```

The following shows the output:

```
+--------------------+
|     organization_id|
+--------------------+
|d56acebe-8fdc-47b...|
|8fa6c3d2-71dc-478...|
+--------------------+
```

Step 5: Put it all together

Now, use AWS Glue to join these relational tables and create one full history table of legislator `memberships` and their corresponding `organizations`.

1. First, join `persons` and `memberships` on `id` and `person_id`.
2. Next, join the result with `orgs` on `org_id` and `organization_id`.
3. Then, drop the redundant fields, `person_id` and `org_id`.

You can do all these operations in one (extended) line of code:

```
l_history = Join.apply(orgs,
                       Join.apply(persons, memberships, 'id', 'person_id'),
                       'org_id', 'organization_id').drop_fields(['person_id',
 'org_id'])
print "Count: ", l_history.count()
l_history.printSchema()
```

The output is as follows:

```
Count:  10439
root
|-- role: string
|-- seats: int
|-- org_name: string
|-- links: array
```

```
|    |-- element: struct
|    |    |-- note: string
|    |    |-- url: string
|-- type: string
|-- sort_name: string
|-- area_id: string
|-- images: array
|    |-- element: struct
|    |    |-- url: string
|-- on_behalf_of_id: string
|-- other_names: array
|    |-- element: struct
|    |    |-- note: string
|    |    |-- name: string
|    |    |-- lang: string
|-- contact_details: array
|    |-- element: struct
|    |    |-- type: string
|    |    |-- value: string
|-- name: string
|-- birth_date: string
|-- organization_id: string
|-- gender: string
|-- classification: string
|-- death_date: string
|-- legislative_period_id: string
|-- identifiers: array
|    |-- element: struct
|    |    |-- scheme: string
|    |    |-- identifier: string
|-- image: string
|-- given_name: string
|-- family_name: string
|-- id: string
|-- start_date: string
|-- end_date: string
```

You now have the final table that you can use for analysis. You can write it out in a compact, efficient format for analytics—namely Parquet—that you can run SQL over in AWS Glue, Amazon Athena, or Amazon Redshift Spectrum.

The following call writes the table across multiple files to support fast parallel reads when doing analysis later:

```
glueContext.write_dynamic_frame.from_options(frame = l_history,
          connection_type = "s3",
          connection_options = {"path": "s3://glue-sample-target/output-dir/
legislator_history"},
          format = "parquet")
```

To put all the history data into a single file, you must convert it to a data frame, repartition it, and write it out:

```
s_history = l_history.toDF().repartition(1)
s_history.write.parquet('s3://glue-sample-target/output-dir/legislator_single')
```

Or, if you want to separate it by the Senate and the House:

```
l_history.toDF().write.parquet('s3://glue-sample-target/output-dir/legislator_part',
                               partitionBy=['org_name'])
```

Step 6: Transform the data for relational databases

AWS Glue makes it easy to write the data to relational databases like Amazon Redshift, even with semi-structured data. It offers a transform `relationalize`, which flattens `DynamicFrames` no matter how complex the objects in the frame might be.

Using the `l_history` `DynamicFrame` in this example, pass in the name of a root table (`hist_root`) and a temporary working path to `relationalize`. This returns a `DynamicFrameCollection`. You can then list the names of the `DynamicFrames` in that collection:

```
dfc = l_history.relationalize("hist_root", "s3://glue-sample-target/temp-dir/")
dfc.keys()
```

The following is the output of the `keys` call:

```
[u'hist_root', u'hist_root_contact_details', u'hist_root_links',
 u'hist_root_other_names', u'hist_root_images', u'hist_root_identifiers']
```

`Relationalize` broke the history table out into six new tables: a root table that contains a record for each object in the `DynamicFrame`, and auxiliary tables for the arrays. Array handling in relational databases is often suboptimal, especially as those arrays become large. Separating the arrays into different tables makes the queries go much faster.

Next, look at the separation by examining `contact_details`:

```
l_history.select_fields('contact_details').printSchema()
dfc.select('hist_root_contact_details').toDF().where("id = 10 or id =
 75").orderBy(['id','index']).show()
```

The following is the output of the `show` call:

```
root
|-- contact_details: array
|    |-- element: struct
|    |    |-- type: string
|    |    |-- value: string
+---+-----+------------------------+-------------------------+
| id|index|contact_details.val.type|contact_details.val.value|
+---+-----+------------------------+-------------------------+
| 10|    0|                     fax|                         |
| 10|    1|                        |             202-225-1314|
| 10|    2|                   phone|                         |
| 10|    3|                        |             202-225-3772|
| 10|    4|                 twitter|                         |
| 10|    5|                        |          MikeRossUpdates|
| 75|    0|                     fax|                         |
| 75|    1|                        |             202-225-7856|
| 75|    2|                   phone|                         |
| 75|    3|                        |             202-225-2711|
| 75|    4|                 twitter|                         |
| 75|    5|                        |                SenCapito|
+---+-----+------------------------+-------------------------+
```

The `contact_details` field was an array of structs in the original `DynamicFrame`. Each element of those arrays is a separate row in the auxiliary table, indexed by `index`. The `id` here is a foreign key into the `hist_root` table with the key `contact_details`:

```
dfc.select('hist_root').toDF().where(
    "contact_details = 10 or contact_details = 75").select(
       ['id', 'given_name', 'family_name', 'contact_details']).show()
```

The following is the output:

```
+--------------------+----------+-----------+---------------+
|                  id|given_name|family_name|contact_details|
+--------------------+----------+-----------+---------------+
|f4fc30ee-7b42-432...|      Mike|       Ross|             10|
|e3c60f34-7d1b-4c0...|   Shelley|     Capito|             75|
+--------------------+----------+-----------+---------------+
```

Notice in these commands that `toDF()` and then a `where` expression are used to filter for the rows that you want to see.

So, joining the `hist_root` table with the auxiliary tables lets you do the following:

- Load data into databases without array support.
- Query each individual item in an array using SQL.

Safely store and access your Amazon Redshift credentials with a AWS Glue connection. For information about how to create your own connection, see Connecting to data.

You are now ready to write your data to a connection by cycling through the `DynamicFrames` one at a time:

```
for df_name in dfc.keys():
  m_df = dfc.select(df_name)
  print "Writing to table: ", df_name
  glueContext.write_dynamic_frame.from_jdbc_conf(frame = m_df, connection settings
 here)
```

Your connection settings will differ based on your type of relational database:

- For instructions on writing to Amazon Redshift consult the section called “Redshift connections”.

- For other databases, consult Connection types and options for ETL in AWS Glue for Spark.

Conclusion

Overall, AWS Glue is very flexible. It lets you accomplish, in a few lines of code, what normally would take days to write. You can find the entire source-to-target ETL scripts in the Python file `join_and_relationalize.py` in the AWS Glue samples on GitHub.

Code example: Data preparation using ResolveChoice, Lambda, and ApplyMapping

The dataset that is used in this example consists of Medicare Provider payment data that was downloaded from two Data.CMS.gov data sets: "Inpatient Prospective Payment System Provider Summary for the Top 100 Diagnosis-Related Groups - FY2011" and "Inpatient Charge Data FY 2011". After downloading the data, we modified the dataset to introduce a couple of erroneous records at the end of the file. This modified file is located in a public Amazon S3 bucket at `s3://awsglue-datasets/examples/medicare/Medicare_Hospital_Provider.csv`.

You can find the source code for this example in the `data_cleaning_and_lambda.py` file in the AWS Glue examples GitHub repository.

The preferred way to debug Python or PySpark scripts while running on AWS is to use Notebooks on AWS Glue Studio.

Step 1: Crawl the data in the Amazon S3 bucket

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. Following the process described in Configuring a crawler, create a new crawler that can crawl the `s3://awsglue-datasets/examples/medicare/Medicare_Hospital_Provider.csv` file, and can place the resulting metadata into a database named `payments` in the AWS Glue Data Catalog.
3. Run the new crawler, and then check the `payments` database. You should find that the crawler has created a metadata table named `medicare` in the database after reading the beginning of the file to determine its format and delimiter.

   The schema of the new `medicare` table is as follows:

   ```
   Column   name                                    Data type
   ==================================================
   ```

```
drg definition                            string
provider id                               bigint
provider name                             string
provider street address                   string
provider city                             string
provider state                            string
provider zip code                         bigint
hospital referral region description      string
total discharges                          bigint
average covered charges                   string
average total payments                    string
average medicare payments                 string
```

Step 2: Add boilerplate script to the development endpoint notebook

Paste the following boilerplate script into the development endpoint notebook to import the AWS Glue libraries that you need, and set up a single `GlueContext`:

```
import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job

glueContext = GlueContext(SparkContext.getOrCreate())
```

Step 3: Compare different schema parsings

Next, you can see if the schema that was recognized by an Apache Spark `DataFrame` is the same as the one that your AWS Glue crawler recorded. Run this code:

```


medicare = spark.read.format(
   "com.databricks.spark.csv").option(
   "header", "true").option(
   "inferSchema", "true").load(
   's3://awsglue-datasets/examples/medicare/Medicare_Hospital_Provider.csv')
medicare.printSchema()
```

Here's the output from the `printSchema` call:

```
root
 |-- DRG Definition: string (nullable = true)
 |-- Provider Id: string (nullable = true)
 |-- Provider Name: string (nullable = true)
 |-- Provider Street Address: string (nullable = true)
 |-- Provider City: string (nullable = true)
 |-- Provider State: string (nullable = true)
 |-- Provider Zip Code: integer (nullable = true)
 |-- Hospital Referral Region Description: string (nullable = true)
 |--  Total Discharges : integer (nullable = true)
 |--  Average Covered Charges : string (nullable = true)
 |--  Average Total Payments : string (nullable = true)
 |-- Average Medicare Payments: string (nullable = true)
```

Next, look at the schema that an AWS Glue `DynamicFrame` generates:

```
medicare_dynamicframe = glueContext.create_dynamic_frame.from_catalog(
       database = "payments",
       table_name = "medicare")
medicare_dynamicframe.printSchema()
```

The output from `printSchema` is as follows:

```
root
 |-- drg definition: string
 |-- provider id: choice
 |    |-- long
 |    |-- string
 |-- provider name: string
 |-- provider street address: string
 |-- provider city: string
 |-- provider state: string
 |-- provider zip code: long
 |-- hospital referral region description: string
 |-- total discharges: long
 |-- average covered charges: string
 |-- average total payments: string
 |-- average medicare payments: string
```

The `DynamicFrame` generates a schema in which `provider id` could be either a `long` or a `string` type. The `DataFrame` schema lists `Provider Id` as being a `string` type, and the Data Catalog lists `provider id` as being a `bigint` type.

Which one is correct? There are two records at the end of the file (out of 160,000 records) with `string` values in that column. These are the erroneous records that were introduced to illustrate a problem.

To address this kind of problem, the AWS Glue `DynamicFrame` introduces the concept of a *choice* type. In this case, the `DynamicFrame` shows that both `long` and `string` values can appear in that column. The AWS Glue crawler missed the `string` values because it considered only a 2 MB prefix of the data. The Apache Spark `DataFrame` considered the whole dataset, but it was forced to assign the most general type to the column, namely `string`. In fact, Spark often resorts to the most general case when there are complex types or variations with which it is unfamiliar.

To query the `provider id` column, resolve the choice type first. You can use the `resolveChoice` transform method in your `DynamicFrame` to convert those `string` values to `long` values with a `cast:long` option:

```
medicare_res = medicare_dynamicframe.resolveChoice(specs = [('provider
 id','cast:long')])
medicare_res.printSchema()
```

The `printSchema` output is now:

```
root
 |-- drg definition: string
 |-- provider id: long
 |-- provider name: string
 |-- provider street address: string
 |-- provider city: string
 |-- provider state: string
 |-- provider zip code: long
 |-- hospital referral region description: string
 |-- total discharges: long
 |-- average covered charges: string
 |-- average total payments: string
 |-- average medicare payments: string
```

Where the value was a `string` that could not be cast, AWS Glue inserted a `null`.

Another option is to convert the choice type to a `struct`, which keeps values of both types.

Next, look at the rows that were anomalous:

```
medicare_res.toDF().where("'provider id' is NULL").show()
```

You see the following:

```
+--------------------+-----------+---------------+-----------------------+-------------
+--------------+-----------------+------------------------------------+----------------
+-----------------------+----------------------+-------------------------+
|      drg definition|provider id|  provider name|provider street address|provider
 city|provider state|provider zip code|hospital referral region description|total
 discharges|average covered charges|average total payments|average medicare payments|
+--------------------+-----------+---------------+-----------------------+-------------
+--------------+-----------------+------------------------------------+----------------
+-----------------------+----------------------+-------------------------+
|948 - SIGNS & SYM...|       null|            INC|       1050 DIVISION ST|
 MAUSTON|            WI|            53948|                        WI - Madison|
      12|              $11961.41|              $4619.00|                 $3775.33|
|948 - SIGNS & SYM...|       null| INC- ST JOSEPH|     5000 W CHAMBERS ST|
 MILWAUKEE|            WI|            53210|                      WI - Milwaukee|
        14|              $10514.28|              $5562.50|                 $4522.78|
+--------------------+-----------+---------------+-----------------------+-------------
+--------------+-----------------+------------------------------------+----------------
+-----------------------+----------------------+-------------------------+
```

Now remove the two malformed records, as follows:

```
medicare_dataframe = medicare_res.toDF()
medicare_dataframe = medicare_dataframe.where("'provider id' is NOT NULL")
```

Step 4: Map the data and use Apache Spark Lambda functions

AWS Glue does not yet directly support Lambda functions, also known as user-defined functions. But you can always convert a `DynamicFrame` to and from an Apache Spark `DataFrame` to take advantage of Spark functionality in addition to the special features of `DynamicFrames`.

Next, turn the payment information into numbers, so analytic engines like Amazon Redshift or Amazon Athena can do their number crunching faster:

```
from pyspark.sql.functions import udf
from pyspark.sql.types import StringType

chop_f = udf(lambda x: x[1:], StringType())
medicare_dataframe = medicare_dataframe.withColumn(
        "ACC", chop_f(
            medicare_dataframe["average covered charges"])).withColumn(
                "ATP", chop_f(
                    medicare_dataframe["average total payments"])).withColumn(
                        "AMP", chop_f(
                            medicare_dataframe["average medicare payments"]))
medicare_dataframe.select(['ACC', 'ATP', 'AMP']).show()
```

The output from the `show` call is as follows:

```
+--------+-------+-------+
|     ACC|    ATP|    AMP|
+--------+-------+-------+
|32963.07|5777.24|4763.73|
|15131.85|5787.57|4976.71|
|37560.37|5434.95|4453.79|
|13998.28|5417.56|4129.16|
|31633.27|5658.33|4851.44|
|16920.79|6653.80|5374.14|
|11977.13|5834.74|4761.41|
|35841.09|8031.12|5858.50|
|28523.39|6113.38|5228.40|
|75233.38|5541.05|4386.94|
|67327.92|5461.57|4493.57|
|39607.28|5356.28|4408.20|
|22862.23|5374.65|4186.02|
|31110.85|5366.23|4376.23|
|25411.33|5282.93|4383.73|
| 9234.51|5676.55|4509.11|
|15895.85|5930.11|3972.85|
|19721.16|6192.54|5179.38|
|10710.88|4968.00|3898.88|
|51343.75|5996.00|4962.45|
+--------+-------+-------+
only showing top 20 rows
```

These are all still strings in the data. We can use the powerful `apply_mapping` transform method to drop, rename, cast, and nest the data so that other data programming languages and systems can easily access it:

```
from awsglue.dynamicframe import DynamicFrame
medicare_tmp_dyf = DynamicFrame.fromDF(medicare_dataframe, glueContext, "nested")
medicare_nest_dyf = medicare_tmp_dyf.apply_mapping([('drg definition', 'string', 'drg',
 'string'),
                 ('provider id', 'long', 'provider.id', 'long'),
                 ('provider name', 'string', 'provider.name', 'string'),
                 ('provider city', 'string', 'provider.city', 'string'),
                 ('provider state', 'string', 'provider.state', 'string'),
                 ('provider zip code', 'long', 'provider.zip', 'long'),
                 ('hospital referral region description', 'string','rr', 'string'),
                 ('ACC', 'string', 'charges.covered', 'double'),
                 ('ATP', 'string', 'charges.total_pay', 'double'),
                 ('AMP', 'string', 'charges.medicare_pay', 'double')])
medicare_nest_dyf.printSchema()
```

The `printSchema` output is as follows:

```
root
 |-- drg: string
 |-- provider: struct
 |    |-- id: long
 |    |-- name: string
 |    |-- city: string
 |    |-- state: string
 |    |-- zip: long
 |-- rr: string
 |-- charges: struct
 |    |-- covered: double
 |    |-- total_pay: double
 |    |-- medicare_pay: double
```

Turning the data back into a Spark `DataFrame`, you can show what it looks like now:

```
medicare_nest_dyf.toDF().show()
```

The output is as follows:

```
+--------------------+--------------------+---------------+--------------------+
|                 drg|            provider|              rr|             charges|
+--------------------+--------------------+----------------+--------------------+
|039 - EXTRACRANIA...|[10001,SOUTHEAST ...|    AL - Dothan|[32963.07,5777.24...|
|039 - EXTRACRANIA...|[10005,MARSHALL M...|AL - Birmingham|[15131.85,5787.57...|
|039 - EXTRACRANIA...|[10006,ELIZA COFF...|AL - Birmingham|[37560.37,5434.95...|
|039 - EXTRACRANIA...|[10011,ST VINCENT...|AL - Birmingham|[13998.28,5417.56...|
|039 - EXTRACRANIA...|[10016,SHELBY BAP...|AL - Birmingham|[31633.27,5658.33...|
|039 - EXTRACRANIA...|[10023,BAPTIST ME...|AL - Montgomery|[16920.79,6653.8,...|
|039 - EXTRACRANIA...|[10029,EAST ALABA...|AL - Birmingham|[11977.13,5834.74...|
|039 - EXTRACRANIA...|[10033,UNIVERSITY...|AL - Birmingham|[35841.09,8031.12...|
|039 - EXTRACRANIA...|[10039,HUNTSVILLE...|AL - Huntsville|[28523.39,6113.38...|
|039 - EXTRACRANIA...|[10040,GADSDEN RE...|AL - Birmingham|[75233.38,5541.05...|
|039 - EXTRACRANIA...|[10046,RIVERVIEW ...|AL - Birmingham|[67327.92,5461.57...|
|039 - EXTRACRANIA...|[10055,FLOWERS HO...|    AL - Dothan|[39607.28,5356.28...|
|039 - EXTRACRANIA...|[10056,ST VINCENT...|AL - Birmingham|[22862.23,5374.65...|
|039 - EXTRACRANIA...|[10078,NORTHEAST ...|AL - Birmingham|[31110.85,5366.23...|
|039 - EXTRACRANIA...|[10083,SOUTH BALD...|    AL - Mobile|[25411.33,5282.93...|
|039 - EXTRACRANIA...|[10085,DECATUR GE...|AL - Huntsville|[9234.51,5676.55,...|
|039 - EXTRACRANIA...|[10090,PROVIDENCE...|    AL - Mobile|[15895.85,5930.11...|
|039 - EXTRACRANIA...|[10092,D C H REGI...|AL - Tuscaloosa|[19721.16,6192.54...|
|039 - EXTRACRANIA...|[10100,THOMAS HOS...|    AL - Mobile|[10710.88,4968.0,...|
|039 - EXTRACRANIA...|[10103,BAPTIST ME...|AL - Birmingham|[51343.75,5996.0,...|
+--------------------+--------------------+----------------+--------------------+
only showing top 20 rows
```

Step 5: Write the data to Apache Parquet

AWS Glue makes it easy to write the data in a format such as Apache Parquet that relational databases can effectively consume:

```
glueContext.write_dynamic_frame.from_options(
       frame = medicare_nest_dyf,
       connection_type = "s3",
       connection_options = {"path": "s3://glue-sample-target/output-dir/
medicare_parquet"},
       format = "parquet")
```

AWS Glue PySpark extensions reference

AWS Glue has created the following extensions to the PySpark Python dialect.

- Accessing parameters using `getResolvedOptions`

- PySpark extension types
- DynamicFrame class
- DynamicFrameCollection class
- DynamicFrameWriter class
- DynamicFrameReader class
- GlueContext class

Accessing parameters using `getResolvedOptions`

The AWS Glue `getResolvedOptions(args, options)` utility function gives you access to the arguments that are passed to your script when you run a job. To use this function, start by importing it from the AWS Glue `utils` module, along with the `sys` module:

```
import sys
from awsglue.utils import getResolvedOptions
```

`getResolvedOptions(args, options)`

- `args` – The list of arguments contained in `sys.argv`.
- `options` – A Python array of the argument names that you want to retrieve.

Example Retrieving arguments passed to a JobRun

Suppose that you created a JobRun in a script, perhaps within a Lambda function:

```
response = client.start_job_run(
             JobName = 'my_test_Job',
             Arguments = {
               '--day_partition_key':   'partition_0',
               '--hour_partition_key':  'partition_1',
               '--day_partition_value':  day_partition_value,
               '--hour_partition_value': hour_partition_value } )
```

To retrieve the arguments that are passed, you can use the `getResolvedOptions` function as follows:

```
import sys
```

```
from awsglue.utils import getResolvedOptions

args = getResolvedOptions(sys.argv,
                          ['JOB_NAME',
                           'day_partition_key',
                           'hour_partition_key',
                           'day_partition_value',
                           'hour_partition_value'])
print "The day-partition key is: ", args['day_partition_key']
print "and the day-partition value is: ", args['day_partition_value']
```

Note that each of the arguments are defined as beginning with two hyphens, then referenced in the script without the hyphens. The arguments use only underscores, not hyphens. Your arguments need to follow this convention to be resolved.

PySpark extension types

The types that are used by the AWS Glue PySpark extensions.

DataType

The base class for the other AWS Glue types.

`__init__(properties={})`

- `properties` – Properties of the data type (optional).

`typeName(cls)`

Returns the type of the AWS Glue type class (that is, the class name with "Type" removed from the end).

- `cls` – An AWS Glue class instance derived from `DataType`.

`jsonValue( )`

Returns a JSON object that contains the data type and properties of the class:

```
  {
    "dataType": typeName,
    "properties": properties
  }
```

AtomicType and simple derivatives

Inherits from and extends the DataType class, and serves as the base class for all the AWS Glue atomic data types.

`fromJsonValue(cls, json_value)`

Initializes a class instance with values from a JSON object.

- `cls` – An AWS Glue type class instance to initialize.
- `json_value` – The JSON object to load key-value pairs from.

The following types are simple derivatives of the AtomicType class:

- `BinaryType` – Binary data.
- `BooleanType` – Boolean values.
- `ByteType` – A byte value.
- `DateType` – A datetime value.
- `DoubleType` – A floating-point double value.
- `IntegerType` – An integer value.
- `LongType` – A long integer value.
- `NullType` – A null value.
- `ShortType` – A short integer value.
- `StringType` – A text string.
- `TimestampType` – A timestamp value (typically in seconds from 1/1/1970).
- `UnknownType` – A value of unidentified type.

DecimalType(AtomicType)

Inherits from and extends the AtomicType class to represent a decimal number (a number expressed in decimal digits, as opposed to binary base-2 numbers).

`__init__(precision=10, scale=2, properties={})`

- `precision` – The number of digits in the decimal number (optional; the default is 10).
- `scale` – The number of digits to the right of the decimal point (optional; the default is 2).
- `properties` – The properties of the decimal number (optional).

EnumType(AtomicType)

Inherits from and extends the AtomicType class to represent an enumeration of valid options.

`__init__(options)`

- `options` – A list of the options being enumerated.

collection types

- ArrayType(DataType)
- ChoiceType(DataType)
- MapType(DataType)
- Field(Object)
- StructType(DataType)
- EntityType(DataType)

ArrayType(DataType)

`__init__(elementType=UnknownType(), properties={})`

- `elementType` – The type of elements in the array (optional; the default is UnknownType).
- `properties` – Properties of the array (optional).

ChoiceType(DataType)

`__init__(choices=[], properties={})`

- `choices` – A list of possible choices (optional).
- `properties` – Properties of these choices (optional).

`add(new_choice)`

Adds a new choice to the list of possible choices.

- `new_choice` – The choice to add to the list of possible choices.

`merge(new_choices)`

Merges a list of new choices with the existing list of choices.

- `new_choices` – A list of new choices to merge with existing choices.

MapType(DataType)

`__init__(valueType=UnknownType, properties={})`

- `valueType` – The type of values in the map (optional; the default is UnknownType).
- `properties` – Properties of the map (optional).

Field(Object)

Creates a field object out of an object that derives from DataType.

`__init__(name, dataType, properties={})`

- `name` – The name to be assigned to the field.
- `dataType` – The object to create a field from.
- `properties` – Properties of the field (optional).

StructType(DataType)

Defines a data structure (`struct`).

`__init__(fields=[], properties={})`

- `fields` – A list of the fields (of type `Field`) to include in the structure (optional).
- `properties` – Properties of the structure (optional).

`add(field)`

- `field` – An object of type `Field` to add to the structure.

`hasField(field)`

Returns `True` if this structure has a field of the same name, or `False` if not.

- `field` – A field name, or an object of type `Field` whose name is used.

`getField(field)`

- `field` – A field name or an object of type `Field` whose name is used. If the structure has a field of the same name, it is returned.

EntityType(DataType)

`__init__(entity, base_type, properties)`

This class is not yet implemented.

other types

- [DataSource(object)]
- [DataSink(object)]

DataSource(object)

`__init__(j_source, sql_ctx, name)`

- `j_source` – The data source.
- `sql_ctx` – The SQL context.
- `name` – The data-source name.

`setFormat(format, **options)`

- `format` – The format to set for the data source.

- `options` – A collection of options to set for the data source. For more information about format options, see the section called “Data format options”.

`getFrame()`

Returns a `DynamicFrame` for the data source.

DataSink(object)

`__init__(j_sink, sql_ctx)`

- `j_sink` – The sink to create.
- `sql_ctx` – The SQL context for the data sink.

`setFormat(format, **options)`

- `format` – The format to set for the data sink.
- `options` – A collection of options to set for the data sink. For more information about format options, see the section called “Data format options”.

`setAccumulableSize(size)`

- `size` – The accumulable size to set, in bytes.

`writeFrame(dynamic_frame, info="")`

- `dynamic_frame` – The `DynamicFrame` to write.
- `info` – Information about the `DynamicFrame` (optional).

`write(dynamic_frame_or_dfc, info="")`

Writes a `DynamicFrame` or a `DynamicFrameCollection`.

- `dynamic_frame_or_dfc` – Either a `DynamicFrame` object or a `DynamicFrameCollection` object to be written.

- `info` – Information about the `DynamicFrame` or `DynamicFrames` to be written (optional).

DynamicFrame class

One of the major abstractions in Apache Spark is the SparkSQL `DataFrame`, which is similar to the `DataFrame` construct found in R and Pandas. A `DataFrame` is similar to a table and supports functional-style (map/reduce/filter/etc.) operations and SQL operations (select, project, aggregate).

`DataFrames` are powerful and widely used, but they have limitations with respect to extract, transform, and load (ETL) operations. Most significantly, they require a schema to be specified before any data is loaded. SparkSQL addresses this by making two passes over the data—the first to infer the schema, and the second to load the data. However, this inference is limited and doesn't address the realities of messy data. For example, the same field might be of a different type in different records. Apache Spark often gives up and reports the type as `string` using the original field text. This might not be correct, and you might want finer control over how schema discrepancies are resolved. And for large datasets, an additional pass over the source data might be prohibitively expensive.

To address these limitations, AWS Glue introduces the `DynamicFrame`. A `DynamicFrame` is similar to a `DataFrame`, except that each record is self-describing, so no schema is required initially. Instead, AWS Glue computes a schema on-the-fly when required, and explicitly encodes schema inconsistencies using a choice (or union) type. You can resolve these inconsistencies to make your datasets compatible with data stores that require a fixed schema.

Similarly, a `DynamicRecord` represents a logical record within a `DynamicFrame`. It is like a row in a Spark `DataFrame`, except that it is self-describing and can be used for data that does not conform to a fixed schema. When using AWS Glue with PySpark, you do not typically manipulate independent `DynamicRecords`. Rather, you will transform the dataset together through its `DynamicFrame`.

You can convert `DynamicFrames` to and from `DataFrames` after you resolve any schema inconsistencies.

— construction —

- __init__
- fromDF
- toDF

__init__

`__init__(jdf, glue_ctx, name)`

- `jdf` – A reference to the data frame in the Java Virtual Machine (JVM).
- `glue_ctx` – A GlueContext class object.
- `name` – An optional name string, empty by default.

fromDF

`fromDF(dataframe, glue_ctx, name)`

Converts a `DataFrame` to a `DynamicFrame` by converting `DataFrame` fields to `DynamicRecord` fields. Returns the new `DynamicFrame`.

A `DynamicRecord` represents a logical record in a `DynamicFrame`. It is similar to a row in a Spark `DataFrame`, except that it is self-describing and can be used for data that does not conform to a fixed schema.

This function expects columns with duplicated names in your `DataFrame` to have already been resolved.

- `dataframe` – The Apache Spark SQL `DataFrame` to convert (required).
- `glue_ctx` – The GlueContext class object that specifies the context for this transform (required).
- `name` – The name of the resulting `DynamicFrame` (optional since AWS Glue 3.0).

toDF

`toDF(options)`

Converts a `DynamicFrame` to an Apache Spark `DataFrame` by converting `DynamicRecords` into `DataFrame` fields. Returns the new `DataFrame`.

A `DynamicRecord` represents a logical record in a `DynamicFrame`. It is similar to a row in a Spark `DataFrame`, except that it is self-describing and can be used for data that does not conform to a fixed schema.

- `options` – A list of options. Allows you to specify additional options for the conversion process. Some valid options that you can use with the `options` parameter:

- `format` – specifies the format of the data, such as json, csv, parquet).
- `separater or sep` – for CSV files, specifies the delimiter.
- `header` – for CSV files, indicates whether the first row is a header (true/false).
- `inferSchema` – directs Spark to infer the schema automatically (true/false).

Here's an example of using the `options` parameter with the `toDF` method:

```
from awsglue.context import GlueContext
from awsglue.dynamicframe import DynamicFrame
from pyspark.context import SparkContext

sc = SparkContext()
glueContext = GlueContext(sc)

csv_dyf = glueContext.create_dynamic_frame.from_options(
  connection_type="s3",
  connection_options={"paths": ["s3://my-bucket/path/to/csv/"]},
  format="csv"
)
csv_cf = csv_dyf.toDF(options={
  "separator": ",",
  "header": "true",
  "inferSchema": "true"
})
```

Specify the target type if you choose the `Project` and `Cast` action type. Examples include the following.

```
>>>toDF([ResolveOption("a.b.c", "KeepAsStruct")])
>>>toDF([ResolveOption("a.b.c", "Project", DoubleType())])
```

— information —

- count
- schema
- printSchema
- show

- repartition
- coalesce

count

`count( )` – Returns the number of rows in the underlying `DataFrame`.

schema

`schema( )` – Returns the schema of this `DynamicFrame`, or if that is not available, the schema of the underlying `DataFrame`.

For more information about the `DynamicFrame` types that make up this schema, see the section called "Types".

printSchema

`printSchema( )` – Prints the schema of the underlying `DataFrame`.

show

`show(num_rows)` – Prints a specified number of rows from the underlying `DataFrame`.

repartition

`repartition(numPartitions)` – Returns a new `DynamicFrame` with `numPartitions` partitions.

coalesce

`coalesce(numPartitions)` – Returns a new `DynamicFrame` with `numPartitions` partitions.

— transforms —

- apply_mapping
- drop_fields
- filter
- join
- map
- mergeDynamicFrame

- relationalize
- rename_field
- resolveChoice
- select_fields
- spigot
- split_fields
- split_rows
- unbox
- the section called “union”
- unnest
- unnest_ddb_json
- write

apply_mapping

`apply_mapping(mappings, transformation_ctx="", info="", stageThreshold=0, totalThreshold=0)`

Applies a declarative mapping to a `DynamicFrame` and returns a new `DynamicFrame` with those mappings applied to the fields that you specify. Unspecified fields are omitted from the new `DynamicFrame`.

- `mappings` – A list of mapping tuples (required). Each consists of: (source column, source type, target column, target type).

  If the source column has a dot "`.`" in the name, you must place backticks "``` `` ```" around it. For example, to map `this.old.name` (string) to `thisNewName`, you would use the following tuple:

  ```
  ("`this.old.name`", "string", "thisNewName", "string")
  ```
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string to be associated with error reporting for this transformation (optional).
- `stageThreshold` – The number of errors encountered during this transformation at which the process should error out (optional). The default is zero, which indicates that the process should not error out.

- `totalThreshold` – The number of errors encountered up to and including this transformation at which the process should error out (optional). The default is zero, which indicates that the process should not error out.

Example: Use apply_mapping to rename fields and change field types

The following code example shows how to use the `apply_mapping` method to rename selected fields and change field types.

> **Note**
>
> To access the dataset that is used in this example, see Code example: Joining and relationalizing data and follow the instructions in Step 1: Crawl the data in the Amazon S3 bucket.

```
# Example: Use apply_mapping to reshape source data into
# the desired column names and types as a new DynamicFrame

from pyspark.context import SparkContext
from awsglue.context import GlueContext

# Create GlueContext
sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)

# Create a DynamicFrame and view its schema
persons = glueContext.create_dynamic_frame.from_catalog(
    database="legislators", table_name="persons_json"
)
print("Schema for the persons DynamicFrame:")
persons.printSchema()

# Select and rename fields, change field type
print("Schema for the persons_mapped DynamicFrame, created with apply_mapping:")
persons_mapped = persons.apply_mapping(
    [
        ("family_name", "String", "last_name", "String"),
        ("name", "String", "first_name", "String"),
        ("birth_date", "String", "date_of_birth", "Date"),
    ]
)
persons_mapped.printSchema()
```

Output

```
Schema for the persons DynamicFrame:
root
|-- family_name: string
|-- name: string
|-- links: array
|    |-- element: struct
|    |    |-- note: string
|    |    |-- url: string
|-- gender: string
|-- image: string
|-- identifiers: array
|    |-- element: struct
|    |    |-- scheme: string
|    |    |-- identifier: string
|-- other_names: array
|    |-- element: struct
|    |    |-- lang: string
|    |    |-- note: string
|    |    |-- name: string
|-- sort_name: string
|-- images: array
|    |-- element: struct
|    |    |-- url: string
|-- given_name: string
|-- birth_date: string
|-- id: string
|-- contact_details: array
|    |-- element: struct
|    |    |-- type: string
|    |    |-- value: string
|-- death_date: string

Schema for the persons_mapped DynamicFrame, created with apply_mapping:
root
|-- last_name: string
|-- first_name: string
|-- date_of_birth: date
```

drop_fields

drop_fields(paths, transformation_ctx="", info="", stageThreshold=0, totalThreshold=0)

Calls the FlatMap class transform to remove fields from a `DynamicFrame`. Returns a new `DynamicFrame` with the specified fields dropped.

- `paths` – A list of strings. Each contains the full path to a field node that you want to drop. You can use dot notation to specify nested fields. For example, if field `first` is a child of field `name` in the tree, you specify `"name.first"` for the path.

  If a field node has a literal `.` in the name, you must enclose the name in backticks (`` ` ``).
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string to be associated with error reporting for this transformation (optional).
- `stageThreshold` – The number of errors encountered during this transformation at which the process should error out (optional). The default is zero, which indicates that the process should not error out.
- `totalThreshold` – The number of errors encountered up to and including this transformation at which the process should error out (optional). The default is zero, which indicates that the process should not error out.

Example: Use drop_fields to remove fields from a DynamicFrame

This code example uses the `drop_fields` method to remove selected top-level and nested fields from a `DynamicFrame`.

Example dataset

The example uses the following dataset that is represented by the `EXAMPLE-FRIENDS-DATA` table in the code:

```
{"name": "Sally", "age": 23, "location": {"state": "WY", "county": "Fremont"},
 "friends": []}
{"name": "Varun", "age": 34, "location": {"state": "NE", "county": "Douglas"},
 "friends": [{"name": "Arjun", "age": 3}]}
```

```
{"name": "George", "age": 52, "location": {"state": "NY"}, "friends": [{"name":
 "Fred"}, {"name": "Amy", "age": 15}]}
{"name": "Haruki", "age": 21, "location": {"state": "AK", "county": "Denali"}}
{"name": "Sheila", "age": 63, "friends": [{"name": "Nancy", "age": 22}]}
```

Example code

```
# Example: Use drop_fields to remove top-level and nested fields from a DynamicFrame.
# Replace MY-EXAMPLE-DATABASE with your Glue Data Catalog database name.
# Replace EXAMPLE-FRIENDS-DATA with your table name.

from pyspark.context import SparkContext
from awsglue.context import GlueContext

# Create GlueContext
sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)

# Create a DynamicFrame from Glue Data Catalog
glue_source_database = "MY-EXAMPLE-DATABASE"
glue_source_table = "EXAMPLE-FRIENDS-DATA"

friends = glueContext.create_dynamic_frame.from_catalog(
    database=glue_source_database, table_name=glue_source_table
)
print("Schema for friends DynamicFrame before calling drop_fields:")
friends.printSchema()

# Remove location.county, remove friends.age, remove age
friends = friends.drop_fields(paths=["age", "location.county", "friends.age"])
print("Schema for friends DynamicFrame after removing age, county, and friend age:")
friends.printSchema()
```

Output

```
Schema for friends DynamicFrame before calling drop_fields:
root
|-- name: string
|-- age: int
|-- location: struct
|    |-- state: string
|    |-- county: string
|-- friends: array
```

```
|    |-- element: struct
|    |    |-- name: string
|    |    |-- age: int

Schema for friends DynamicFrame after removing age, county, and friend age:
root
|-- name: string
|-- location: struct
|    |-- state: string
|-- friends: array
|    |-- element: struct
|    |    |-- name: string
```

filter

`filter(f, transformation_ctx="", info="", stageThreshold=0, totalThreshold=0)`

Returns a new `DynamicFrame` that contains all `DynamicRecords` within the input `DynamicFrame` that satisfy the specified predicate function `f`.

- `f` – The predicate function to apply to the `DynamicFrame`. The function must take a `DynamicRecord` as an argument and return True if the `DynamicRecord` meets the filter requirements, or False if not (required).

  A `DynamicRecord` represents a logical record in a `DynamicFrame`. It's similar to a row in a Spark `DataFrame`, except that it is self-describing and can be used for data that doesn't conform to a fixed schema.
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string to be associated with error reporting for this transformation (optional).
- `stageThreshold` – The number of errors encountered during this transformation at which the process should error out (optional). The default is zero, which indicates that the process should not error out.
- `totalThreshold` – The number of errors encountered up to and including this transformation at which the process should error out (optional). The default is zero, which indicates that the process should not error out.

Example: Use filter to get a filtered selection of fields

This example uses the `filter` method to create a new `DynamicFrame` that includes a filtered selection of another `DynamicFrame`'s fields.

Like the `map` method, `filter` takes a function as an argument that gets applied to each record in the original `DynamicFrame`. The function takes a record as an input and returns a Boolean value. If the return value is true, the record gets included in the resulting `DynamicFrame`. If it's false, the record is left out.

> **Note**
>
> To access the dataset that is used in this example, see Code example: Data preparation using ResolveChoice, Lambda, and ApplyMapping and follow the instructions in Step 1: Crawl the data in the Amazon S3 bucket.

```
# Example: Use filter to create a new DynamicFrame
# with a filtered selection of records

from pyspark.context import SparkContext
from awsglue.context import GlueContext

# Create GlueContext
sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)

# Create DynamicFrame from Glue Data Catalog
medicare = glueContext.create_dynamic_frame.from_options(
    "s3",
    {
        "paths": [
            "s3://awsglue-datasets/examples/medicare/Medicare_Hospital_Provider.csv"
        ]
    },
    "csv",
    {"withHeader": True},
)

# Create filtered DynamicFrame with custom lambda
# to filter records by Provider State and Provider City
sac_or_mon = medicare.filter(
    f=lambda x: x["Provider State"] in ["CA", "AL"]
    and x["Provider City"] in ["SACRAMENTO", "MONTGOMERY"]
)

# Compare record counts
print("Unfiltered record count: ", medicare.count())
print("Filtered record count:  ", sac_or_mon.count())
```

Output

```
Unfiltered record count:  163065
Filtered record count:   564
```

join

join(paths1, paths2, frame2, transformation_ctx="", info="", stageThreshold=0, totalThreshold=0)

Performs an equality join with another `DynamicFrame` and returns the resulting `DynamicFrame`.

- `paths1` – A list of the keys in this frame to join.
- `paths2` – A list of the keys in the other frame to join.
- `frame2` – The other `DynamicFrame` to join.
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string to be associated with error reporting for this transformation (optional).
- `stageThreshold` – The number of errors encountered during this transformation at which the process should error out (optional). The default is zero, which indicates that the process should not error out.
- `totalThreshold` – The number of errors encountered up to and including this transformation at which the process should error out (optional). The default is zero, which indicates that the process should not error out.

Example: Use join to combine `DynamicFrames`

This example uses the `join` method to perform a join on three `DynamicFrames`. AWS Glue performs the join based on the field keys that you provide. The resulting `DynamicFrame` contains rows from the two original frames where the specified keys match.

Note that the `join` transform keeps all fields intact. This means that the fields that you specify to match appear in the resulting DynamicFrame, even if they're redundant and contain the same keys. In this example, we use `drop_fields` to remove these redundant keys after the join.

> **Note**
>
> To access the dataset that is used in this example, see Code example: Joining and relationalizing data and follow the instructions in Step 1: Crawl the data in the Amazon S3 bucket.

```
# Example: Use join to combine data from three DynamicFrames

from pyspark.context import SparkContext
from awsglue.context import GlueContext

# Create GlueContext
sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)

# Load DynamicFrames from Glue Data Catalog
persons = glueContext.create_dynamic_frame.from_catalog(
    database="legislators", table_name="persons_json"
)
memberships = glueContext.create_dynamic_frame.from_catalog(
    database="legislators", table_name="memberships_json"
)
orgs = glueContext.create_dynamic_frame.from_catalog(
    database="legislators", table_name="organizations_json"
)
print("Schema for the persons DynamicFrame:")
persons.printSchema()
print("Schema for the memberships DynamicFrame:")
memberships.printSchema()
print("Schema for the orgs DynamicFrame:")
orgs.printSchema()

# Join persons and memberships by ID
persons_memberships = persons.join(
    paths1=["id"], paths2=["person_id"], frame2=memberships
)
```

```
# Rename and drop fields from orgs
# to prevent field name collisions with persons_memberships
orgs = (
    orgs.drop_fields(["other_names", "identifiers"])
    .rename_field("id", "org_id")
    .rename_field("name", "org_name")
)

# Create final join of all three DynamicFrames
legislators_combined = orgs.join(
    paths1=["org_id"], paths2=["organization_id"], frame2=persons_memberships
).drop_fields(["person_id", "org_id"])

# Inspect the schema for the joined data
print("Schema for the new legislators_combined DynamicFrame:")
legislators_combined.printSchema()
```

Output

```
Schema for the persons DynamicFrame:
root
|-- family_name: string
|-- name: string
|-- links: array
|    |-- element: struct
|    |    |-- note: string
|    |    |-- url: string
|-- gender: string
|-- image: string
|-- identifiers: array
|    |-- element: struct
|    |    |-- scheme: string
|    |    |-- identifier: string
|-- other_names: array
|    |-- element: struct
|    |    |-- lang: string
|    |    |-- note: string
|    |    |-- name: string
|-- sort_name: string
|-- images: array
|    |-- element: struct
|    |    |-- url: string
|-- given_name: string
|-- birth_date: string
|-- id: string
|-- contact_details: array
|    |-- element: struct
|    |    |-- type: string
|    |    |-- value: string
|-- death_date: string

Schema for the memberships DynamicFrame:
root
|-- area_id: string
|-- on_behalf_of_id: string
|-- organization_id: string
|-- role: string
|-- person_id: string
|-- legislative_period_id: string
|-- start_date: string
|-- end_date: string

Schema for the orgs DynamicFrame:
root
|-- identifiers: array
|    |-- element: struct
|    |    |-- scheme: string
|    |    |-- identifier: string
|-- other_names: array
|    |-- element: struct
|    |    |-- lang: string
|    |    |-- note: string
|    |    |-- name: string
|-- id: string
|-- classification: string
|-- name: string
|-- links: array
|    |-- element: struct
|    |    |-- note: string
|    |    |-- url: string
|-- image: string
|-- seats: int
|-- type: string

Schema for the new legislators_combined DynamicFrame:
root
|-- role: string
|-- seats: int
|-- org_name: string
|-- links: array
|    |-- element: struct
|    |    |-- note: string
|    |    |-- url: string
|-- type: string
|-- sort_name: string
|-- area_id: string
|-- images: array
|    |-- element: struct
|    |    |-- url: string
|-- on_behalf_of_id: string
|-- other_names: array
|    |-- element: struct
|    |    |-- note: string
|    |    |-- name: string
|    |    |-- lang: string
|-- contact_details: array
|    |-- element: struct
|    |    |-- type: string
|    |    |-- value: string
|-- name: string
|-- birth_date: string
|-- organization_id: string
|-- gender: string
|-- classification: string
|-- legislative_period_id: string
|-- identifiers: array
|    |-- element: struct
|    |    |-- scheme: string
|    |    |-- identifier: string
|-- image: string
|-- given_name: string
|-- start_date: string
|-- family_name: string
|-- id: string
|-- death_date: string
|-- end_date: string
```

map

`map(f, transformation_ctx="", info="", stageThreshold=0, totalThreshold=0)`

Returns a new `DynamicFrame` that results from applying the specified mapping function to all records in the original `DynamicFrame`.

- `f` – The mapping function to apply to all records in the `DynamicFrame`. The function must take a `DynamicRecord` as an argument and return a new `DynamicRecord` (required).

  A `DynamicRecord` represents a logical record in a `DynamicFrame`. It's similar to a row in an Apache Spark `DataFrame`, except that it is self-describing and can be used for data that doesn't conform to a fixed schema.
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string that is associated with errors in the transformation (optional).
- `stageThreshold` – The maximum number of errors that can occur in the transformation before it errors out (optional). The default is zero.
- `totalThreshold` – The maximum number of errors that can occur overall before processing errors out (optional). The default is zero.

Example: Use map to apply a function to every record in a `DynamicFrame`

This example shows how to use the `map` method to apply a function to every record of a `DynamicFrame`. Specifically, this example applies a function called `MergeAddress` to each record in order to merge several address fields into a single `struct` type.

> **Note**
>
> To access the dataset that is used in this example, see Code example: Data preparation using ResolveChoice, Lambda, and ApplyMapping and follow the instructions in Step 1: Crawl the data in the Amazon S3 bucket.

```
# Example: Use map to combine fields in all records
# of a DynamicFrame

from pyspark.context import SparkContext
from awsglue.context import GlueContext
```

```
# Create GlueContext
sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)

# Create a DynamicFrame and view its schema
medicare = glueContext.create_dynamic_frame.from_options(
        "s3",
        {"paths": ["s3://awsglue-datasets/examples/medicare/
Medicare_Hospital_Provider.csv"]},
        "csv",
        {"withHeader": True})
print("Schema for medicare DynamicFrame:")
medicare.printSchema()

# Define a function to supply to the map transform
# that merges address fields into a single field
def MergeAddress(rec):
  rec["Address"] = {}
  rec["Address"]["Street"] = rec["Provider Street Address"]
  rec["Address"]["City"] = rec["Provider City"]
  rec["Address"]["State"] = rec["Provider State"]
  rec["Address"]["Zip.Code"] = rec["Provider Zip Code"]
  rec["Address"]["Array"] = [rec["Provider Street Address"], rec["Provider City"],
 rec["Provider State"], rec["Provider Zip Code"]]
  del rec["Provider Street Address"]
  del rec["Provider City"]
  del rec["Provider State"]
  del rec["Provider Zip Code"]
  return rec


# Use map to apply MergeAddress to every record
mapped_medicare = medicare.map(f = MergeAddress)
print("Schema for mapped_medicare DynamicFrame:")
mapped_medicare.printSchema()
```

Output

```
Schema for medicare DynamicFrame:
root
|-- DRG Definition: string
|-- Provider Id: string
|-- Provider Name: string
|-- Provider Street Address: string
|-- Provider City: string
|-- Provider State: string
|-- Provider Zip Code: string
|-- Hospital Referral Region Description: string
|-- Total Discharges: string
|-- Average Covered Charges: string
|-- Average Total Payments: string
|-- Average Medicare Payments: string

Schema for mapped_medicare DynamicFrame:
root
|-- Average Total Payments: string
|-- Average Covered Charges: string
|-- DRG Definition: string
|-- Average Medicare Payments: string
|-- Hospital Referral Region Description: string
|-- Address: struct
|    |-- Zip.Code: string
|    |-- City: string
|    |-- Array: array
|    |    |-- element: string
|    |-- State: string
|    |-- Street: string
|-- Provider Id: string
|-- Total Discharges: string
|-- Provider Name: string
```

mergeDynamicFrame

`mergeDynamicFrame(stage_dynamic_frame, primary_keys, transformation_ctx = "", options = {}, info = "", stageThreshold = 0, totalThreshold = 0)`

Merges this `DynamicFrame` with a staging `DynamicFrame` based on the specified primary keys to identify records. Duplicate records (records with the same primary keys) are not deduplicated. If there is no matching record in the staging frame, all records (including duplicates) are retained from the source. If the staging frame has matching records, the records from the staging frame overwrite the records in the source in AWS Glue.

- `stage_dynamic_frame` – The staging `DynamicFrame` to merge.

- `primary_keys` – The list of primary key fields to match records from the source and staging dynamic frames.
- `transformation_ctx` – A unique string that is used to retrieve metadata about the current transformation (optional).
- `options` – A string of JSON name-value pairs that provide additional information for this transformation. This argument is not currently used.
- `info` – A `String`. Any string to be associated with errors in this transformation.
- `stageThreshold` – A `Long`. The number of errors in the given transformation for which the processing needs to error out.
- `totalThreshold` – A `Long`. The total number of errors up to and including this transformation for which the processing needs to error out.

This method returns a new `DynamicFrame` that is obtained by merging this `DynamicFrame` with the staging `DynamicFrame`.

The returned `DynamicFrame` contains record A in these cases:

- If `A` exists in both the source frame and the staging frame, then `A` in the staging frame is returned.
- If `A` is in the source table and `A.primaryKeys` is not in the `stagingDynamicFrame`, `A` is not updated in the staging table.

The source frame and staging frame don't need to have the same schema.

Example: Use mergeDynamicFrame to merge two `DynamicFrames` based on a primary key

The following code example shows how to use the `mergeDynamicFrame` method to merge a `DynamicFrame` with a "staging" `DynamicFrame`, based on the primary key `id`.

Example dataset

The example uses two `DynamicFrames` from a `DynamicFrameCollection` called `split_rows_collection`. The following is the list of keys in `split_rows_collection`.

```
dict_keys(['high', 'low'])
```

Example code

```
# Example: Use mergeDynamicFrame to merge DynamicFrames
# based on a set of specified primary keys

from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.transforms import SelectFromCollection

# Inspect the original DynamicFrames
frame_low = SelectFromCollection.apply(dfc=split_rows_collection, key="low")
print("Inspect the DynamicFrame that contains rows where ID < 10")
frame_low.toDF().show()

frame_high = SelectFromCollection.apply(dfc=split_rows_collection, key="high")
print("Inspect the DynamicFrame that contains rows where ID > 10")
frame_high.toDF().show()

# Merge the DynamicFrames based on the "id" primary key
merged_high_low = frame_high.mergeDynamicFrame(
    stage_dynamic_frame=frame_low, primary_keys=["id"]
)

# View the results where the ID is 1 or 20
print("Inspect the merged DynamicFrame that contains the combined rows")
merged_high_low.toDF().where("id = 1 or id= 20").orderBy("id").show()
```

Output

```
Inspect the DynamicFrame that contains rows where ID < 10
+---+-----+------------------------+-------------------------+
| id|index|contact_details.val.type|contact_details.val.value|
+---+-----+------------------------+-------------------------+
|  1|    0|                     fax|             202-225-3307|
|  1|    1|                   phone|             202-225-5731|
|  2|    0|                     fax|             202-225-3307|
|  2|    1|                   phone|             202-225-5731|
|  3|    0|                     fax|             202-225-3307|
|  3|    1|                   phone|             202-225-5731|
|  4|    0|                     fax|             202-225-3307|
|  4|    1|                   phone|             202-225-5731|
|  5|    0|                     fax|             202-225-3307|
|  5|    1|                   phone|             202-225-5731|
|  6|    0|                     fax|             202-225-3307|
|  6|    1|                   phone|             202-225-5731|
|  7|    0|                     fax|             202-225-3307|
|  7|    1|                   phone|             202-225-5731|
|  8|    0|                     fax|             202-225-3307|
|  8|    1|                   phone|             202-225-5731|
|  9|    0|                     fax|             202-225-3307|
|  9|    1|                   phone|             202-225-5731|
| 10|    0|                     fax|             202-225-6328|
| 10|    1|                   phone|             202-225-4576|
+---+-----+------------------------+-------------------------+
only showing top 20 rows

Inspect the DynamicFrame that contains rows where ID > 10
+---+-----+------------------------+-------------------------+
| id|index|contact_details.val.type|contact_details.val.value|
+---+-----+------------------------+-------------------------+
| 11|    0|                     fax|             202-225-6328|
| 11|    1|                   phone|             202-225-4576|
| 11|    2|                 twitter|           RepTrentFranks|
| 12|    0|                     fax|             202-225-6328|
| 12|    1|                   phone|             202-225-4576|
| 12|    2|                 twitter|           RepTrentFranks|
| 13|    0|                     fax|             202-225-6328|
| 13|    1|                   phone|             202-225-4576|
| 13|    2|                 twitter|           RepTrentFranks|
| 14|    0|                     fax|             202-225-6328|
| 14|    1|                   phone|             202-225-4576|
| 14|    2|                 twitter|           RepTrentFranks|
| 15|    0|                     fax|             202-225-6328|
| 15|    1|                   phone|             202-225-4576|
| 15|    2|                 twitter|           RepTrentFranks|
| 16|    0|                     fax|             202-225-6328|
| 16|    1|                   phone|             202-225-4576|
| 16|    2|                 twitter|           RepTrentFranks|
| 17|    0|                     fax|             202-225-6328|
| 17|    1|                   phone|             202-225-4576|
+---+-----+------------------------+-------------------------+
only showing top 20 rows

Inspect the merged DynamicFrame that contains the combined rows
+---+-----+------------------------+-------------------------+
| id|index|contact_details.val.type|contact_details.val.value|
+---+-----+------------------------+-------------------------+
|  1|    0|                     fax|             202-225-3307|
|  1|    1|                   phone|             202-225-5731|
| 20|    0|                     fax|                202-225-5604|
| 20|    1|                   phone|                202-225-6536|
| 20|    2|                 twitter|                   USRepLong|
+---+-----+------------------------+-------------------------+
```

relationalize

`relationalize(root_table_name, staging_path, options, transformation_ctx="", info="", stageThreshold=0, totalThreshold=0)`

Converts a `DynamicFrame` into a form that fits within a relational database. Relationalizing a `DynamicFrame` is especially useful when you want to move data from a NoSQL environment like DynamoDB into a relational database like MySQL.

The transform generates a list of frames by unnesting nested columns and pivoting array columns. You can join the pivoted array columns to the root table by using the join key that is generated during the unnest phase.

- `root_table_name` – The name for the root table.
- `staging_path` – The path where the method can store partitions of pivoted tables in CSV format (optional). Pivoted tables are read back from this path.
- `options` – A dictionary of optional parameters.
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string to be associated with error reporting for this transformation (optional).
- `stageThreshold` – The number of errors encountered during this transformation at which the process should error out (optional). The default is zero, which indicates that the process should not error out.
- `totalThreshold` – The number of errors encountered up to and including this transformation at which the process should error out (optional). The default is zero, which indicates that the process should not error out.

Example: Use relationalize to flatten a nested schema in a `DynamicFrame`

This code example uses the `relationalize` method to flatten a nested schema into a form that fits into a relational database.

Example dataset

The example uses a `DynamicFrame` called `legislators_combined` with the following schema. `legislators_combined` has multiple nested fields such as `links`, `images`, and `contact_details`, which will be flattened by the `relationalize` transform.

```
root
|-- role: string
|-- seats: int
|-- org_name: string
|-- links: array
|    |-- element: struct
|    |    |-- note: string
|    |    |-- url: string
|-- type: string
|-- sort_name: string
|-- area_id: string
|-- images: array
|    |-- element: struct
|    |    |-- url: string
|-- on_behalf_of_id: string
|-- other_names: array
|    |-- element: struct
|    |    |-- note: string
|    |    |-- name: string
|    |    |-- lang: string
|-- contact_details: array
|    |-- element: struct
|    |    |-- type: string
|    |    |-- value: string
|-- name: string
|-- birth_date: string
|-- organization_id: string
|-- gender: string
|-- classification: string
|-- legislative_period_id: string
|-- identifiers: array
|    |-- element: struct
|    |    |-- scheme: string
|    |    |-- identifier: string
|-- image: string
|-- given_name: string
|-- start_date: string
|-- family_name: string
|-- id: string
|-- death_date: string
|-- end_date: string
```

Example code

```
# Example: Use relationalize to flatten
# a nested schema into a format that fits
# into a relational database.
# Replace DOC-EXAMPLE-S3-BUCKET/tmpDir with your own location.

from pyspark.context import SparkContext
from awsglue.context import GlueContext

# Create GlueContext
sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)

# Apply relationalize and inspect new tables
legislators_relationalized = legislators_combined.relationalize(
    "l_root", "s3://DOC-EXAMPLE-BUCKET/tmpDir"
)
legislators_relationalized.keys()

# Compare the schema of the contact_details
# nested field to the new relationalized table that
# represents it
legislators_combined.select_fields("contact_details").printSchema()
legislators_relationalized.select("l_root_contact_details").toDF().where(
    "id = 10 or id = 75"
).orderBy(["id", "index"]).show()
```

Output

The following output lets you compare the schema of the nested field called `contact_details` to the table that the `relationalize` transform created. Notice that the table records link back to the main table using a foreign key called `id` and an `index` column that represents the positions of the array.

```
dict_keys(['l_root', 'l_root_images', 'l_root_links', 'l_root_other_names',
 'l_root_contact_details', 'l_root_identifiers'])

root
```

```
|-- contact_details: array
|    |-- element: struct
|    |    |-- type: string
|    |    |-- value: string

+---+-----+-----------------------+------------------------+
| id|index|contact_details.val.type|contact_details.val.value|
+---+-----+-----------------------+------------------------+
| 10|    0|                    fax|            202-225-4160|
| 10|    1|                  phone|            202-225-3436|
| 75|    0|                    fax|            202-225-6791|
| 75|    1|                  phone|            202-225-2861|
| 75|    2|                twitter|              RepSamFarr|
+---+-----+-----------------------+------------------------+
```

rename_field

`rename_field(oldName, newName, transformation_ctx="", info="", stageThreshold=0, totalThreshold=0)`

Renames a field in this `DynamicFrame` and returns a new `DynamicFrame` with the field renamed.

- `oldName` – The full path to the node you want to rename.

  If the old name has dots in it, `RenameField` doesn't work unless you place backticks around it (`` ` ``). For example, to replace `this.old.name` with `thisNewName`, you would call rename_field as follows.

  ```
  newDyF = oldDyF.rename_field("`this.old.name`", "thisNewName")
  ```

- `newName` – The new name, as a full path.
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string to be associated with error reporting for this transformation (optional).
- `stageThreshold` – The number of errors encountered during this transformation at which the process should error out (optional). The default is zero, which indicates that the process should not error out.
- `totalThreshold` – The number of errors encountered up to and including this transformation at which the process should error out (optional). The default is zero, which indicates that the process should not error out.

Example: Use rename_field to rename fields in a `DynamicFrame`

This code example uses the `rename_field` method to rename fields in a `DynamicFrame`. Notice that the example uses method chaining to rename multiple fields at the same time.

> **Note**
>
> To access the dataset that is used in this example, see Code example: Joining and relationalizing data and follow the instructions in Step 1: Crawl the data in the Amazon S3 bucket.

Example code

```
# Example: Use rename_field to rename fields
# in a DynamicFrame

from pyspark.context import SparkContext
from awsglue.context import GlueContext

# Create GlueContext
sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)

# Inspect the original orgs schema
orgs = glueContext.create_dynamic_frame.from_catalog(
    database="legislators", table_name="organizations_json"
)
print("Original orgs schema: ")
orgs.printSchema()

# Rename fields and view the new schema
orgs = orgs.rename_field("id", "org_id").rename_field("name", "org_name")
print("New orgs schema with renamed fields: ")
orgs.printSchema()
```

Output

```
Original orgs schema:
root
|-- identifiers: array
|    |-- element: struct
|    |    |-- scheme: string
|    |    |-- identifier: string
|-- other_names: array
|    |-- element: struct
|    |    |-- lang: string
|    |    |-- note: string
|    |    |-- name: string
|-- id: string
|-- classification: string
|-- name: string
|-- links: array
|    |-- element: struct
|    |    |-- note: string
|    |    |-- url: string
|-- image: string
|-- seats: int
|-- type: string

New orgs schema with renamed fields:
root
|-- identifiers: array
|    |-- element: struct
|    |    |-- scheme: string
|    |    |-- identifier: string
|-- other_names: array
|    |-- element: struct
|    |    |-- lang: string
|    |    |-- note: string
|    |    |-- name: string
|-- classification: string
|-- org_id: string
|-- org_name: string
|-- links: array
|    |-- element: struct
|    |    |-- note: string
|    |    |-- url: string
|-- image: string
|-- seats: int
|-- type: string
```

resolveChoice

`resolveChoice(specs = None, choice = "" , database = None , table_name = None , transformation_ctx="", info="", stageThreshold=0, totalThreshold=0, catalog_id = None)`

Resolves a choice type within this `DynamicFrame` and returns the new `DynamicFrame`.

- `specs` – A list of specific ambiguities to resolve, each in the form of a tuple: `(field_path, action)`.

  There are two ways to use `resolveChoice`. The first is to use the `specs` argument to specify a sequence of specific fields and how to resolve them. The other mode for `resolveChoice` is to use the `choice` argument to specify a single resolution for all `ChoiceTypes`.

  Values for `specs` are specified as tuples made up of `(field_path, action)` pairs. The `field_path` value identifies a specific ambiguous element, and the `action` value identifies the corresponding resolution. The following are the possible actions:

  - `cast:`*`type`* – Attempts to cast all values to the specified type. For example: `cast:int`.
  - `make_cols` – Converts each distinct type to a column with the name *`columnName_type`*. It resolves a potential ambiguity by flattening the data. For example, if `columnA` could be an `int` or a `string`, the resolution would be to produce two columns named `columnA_int` and `columnA_string` in the resulting `DynamicFrame`.
  - `make_struct` – Resolves a potential ambiguity by using a `struct` to represent the data. For example, if data in a column could be an `int` or a `string`, the `make_struct` action produces a column of structures in the resulting `DynamicFrame`. Each structure contains both an `int` and a `string`.
  - `project:`*`type`* – Resolves a potential ambiguity by projecting all the data to one of the possible data types. For example, if data in a column could be an `int` or a `string`, using a `project:string` action produces a column in the resulting `DynamicFrame` where all the `int` values have been converted to strings.

  If the `field_path` identifies an array, place empty square brackets after the name of the array to avoid ambiguity. For example, suppose you are working with data structured as follows:

```
"myList": [
  { "price": 100.00 },
  { "price": "$100.00" }
]
```

You can select the numeric rather than the string version of the price by setting the `field_path` to `"myList[].price"`, and setting the `action` to `"cast:double"`.

> **Note**
>
> You can only use one of the `specs` and `choice` parameters. If the `specs` parameter is not `None`, then the `choice` parameter must be an empty string. Conversely, if the `choice` is not an empty string, then the `specs` parameter must be `None`.

- `choice` – Specifies a single resolution for all `ChoiceTypes`. You can use this in cases where the complete list of `ChoiceTypes` is unknown before runtime. In addition to the actions listed previously for `specs`, this argument also supports the following action:
  - `match_catalog` – Attempts to cast each `ChoiceType` to the corresponding type in the specified Data Catalog table.
- `database` – The Data Catalog database to use with the `match_catalog` action.
- `table_name` – The Data Catalog table to use with the `match_catalog` action.
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string to be associated with error reporting for this transformation (optional).
- `stageThreshold` – The number of errors encountered during this transformation at which the process should error out (optional). The default is zero, which indicates that the process should not error out.
- `totalThreshold` – The number of errors encountered up to and including this transformation at which the process should error out (optional).The default is zero, which indicates that the process should not error out.
- `catalog_id` – The catalog ID of the Data Catalog being accessed (the account ID of the Data Catalog). When set to `None` (default value), it uses the catalog ID of the calling account.

Example: Use resolveChoice to handle a column that contains multiple types

This code example uses the `resolveChoice` method to specify how to handle a `DynamicFrame` column that contains values of multiple types. The example demonstrates two common ways to handle a column with different types:

- Cast the column to a single data type.

- Retain all types in separate columns.

Example dataset

Note

To access the dataset that is used in this example, see Code example: Data preparation using ResolveChoice, Lambda, and ApplyMapping and follow the instructions in Step 1: Crawl the data in the Amazon S3 bucket.

The example uses a `DynamicFrame` called `medicare` with the following schema:

```
root
|-- drg definition: string
|-- provider id: choice
|    |-- long
|    |-- string
|-- provider name: string
|-- provider street address: string
|-- provider city: string
|-- provider state: string
|-- provider zip code: long
|-- hospital referral region description: string
|-- total discharges: long
|-- average covered charges: string
|-- average total payments: string
|-- average medicare payments: string
```

Example code

```
# Example: Use resolveChoice to handle
# a column that contains multiple types

from pyspark.context import SparkContext
from awsglue.context import GlueContext

# Create GlueContext
sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)
```

```
# Load the input data and inspect the "provider id" column
medicare = glueContext.create_dynamic_frame.from_catalog(
    database="payments", table_name="medicare_hospital_provider_csv"
)
print("Inspect the provider id column:")
medicare.toDF().select("provider id").show()

# Cast provider id to type long
medicare_resolved_long = medicare.resolveChoice(specs=[("provider id", "cast:long")])
print("Schema after casting provider id to type long:")
medicare_resolved_long.printSchema()
medicare_resolved_long.toDF().select("provider id").show()

# Create separate columns
# for each provider id type
medicare_resolved_cols = medicare.resolveChoice(choice="make_cols")
print("Schema after creating separate columns for each type:")
medicare_resolved_cols.printSchema()
medicare_resolved_cols.toDF().select("provider id_long", "provider id_string").show()
```

Output

```
Inspect the 'provider id' column:
+-----------+
|provider id|
+-----------+
|   [10001,]|
|   [10005,]|
|   [10006,]|
|   [10011,]|
|   [10016,]|
|   [10023,]|
|   [10029,]|
|   [10033,]|
|   [10039,]|
|   [10040,]|
|   [10046,]|
|   [10055,]|
|   [10056,]|
|   [10078,]|
|   [10083,]|
|   [10085,]|
|   [10090,]|
|   [10092,]|
|   [10100,]|
|   [10103,]|
+-----------+
only showing top 20 rows

Schema after casting 'provider id' to type long:
root
|-- drg definition: string
|-- provider id: long
|-- provider name: string
|-- provider street address: string
|-- provider city: string
|-- provider state: string
|-- provider zip code: long
|-- hospital referral region description: string
|-- total discharges: long
|-- average covered charges: string
|-- average total payments: string
|-- average medicare payments: string

+-----------+
|provider id|
+-----------+
|      10001|
|      10005|
|      10006|
|      10011|
|      10016|
|      10023|
|      10029|
|      10033|
|      10039|
|      10040|
|      10046|
|      10055|
|      10056|
|      10078|
|      10083|
|      10085|
|      10090|
|      10092|
|      10100|
|      10103|
+-----------+
only showing top 20 rows

Schema after creating separate columns for each type:
root
|-- drg definition: string
|-- provider id_string: string
|-- provider id_long: long
|-- provider name: string
|-- provider street address: string
|-- provider city: string
|-- provider state: string
|-- provider zip code: long
|-- hospital referral region description: string
|-- total discharges: long
|-- average covered charges: string
|-- average total payments: string
|-- average medicare payments: string

+----------------+------------------+
|provider id_long|provider id_string|
+----------------+------------------+
|           10001|              null|
|           10005|              null|
|           10006|              null|
|           10011|              null|
|           10016|              null|
|           10023|              null|
|           10029|              null|
|           10033|              null|
|           10039|              null|
|           10040|              null|
|           10046|              null|
|           10055|              null|
|           10056|              null|
|           10078|              null|
|           10083|              null|
|           10085|              null|
|           10090|              null|
|           10092|              null|
|           10100|              null|
|           10103|              null|
+----------------+------------------+
only showing top 20 rows
```

select_fields

`select_fields(paths, transformation_ctx="", info="", stageThreshold=0, totalThreshold=0)`

Returns a new `DynamicFrame` that contains the selected fields.

- `paths` – A list of strings. Each string is a path to a top-level node that you want to select.
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string to be associated with error reporting for this transformation (optional).
- `stageThreshold` – The number of errors encountered during this transformation at which the process should error out (optional). The default is zero, which indicates that the process should not error out.
- `totalThreshold` – The number of errors encountered up to and including this transformation at which the process should error out (optional). The default is zero, which indicates that the process should not error out.

Example: Use select_fields to create a new `DynamicFrame` with chosen fields

The following code example shows how to use the `select_fields` method to create a new `DynamicFrame` with a chosen list of fields from an existing `DynamicFrame`.

> **Note**
>
> To access the dataset that is used in this example, see Code example: Joining and relationalizing data and follow the instructions in Step 1: Crawl the data in the Amazon S3 bucket.

```
# Example: Use select_fields to select specific fields from a DynamicFrame

from pyspark.context import SparkContext
from awsglue.context import GlueContext

# Create GlueContext
sc = SparkContext.getOrCreate()
```

```
glueContext = GlueContext(sc)

# Create a DynamicFrame and view its schema
persons = glueContext.create_dynamic_frame.from_catalog(
    database="legislators", table_name="persons_json"
)
print("Schema for the persons DynamicFrame:")
persons.printSchema()

# Create a new DynamicFrame with chosen fields
names = persons.select_fields(paths=["family_name", "given_name"])
print("Schema for the names DynamicFrame, created with select_fields:")
names.printSchema()
names.toDF().show()
```

Output

```
Schema for the persons DynamicFrame:
root
|-- family_name: string
|-- name: string
|-- links: array
|    |-- element: struct
|    |    |-- note: string
|    |    |-- url: string
|-- gender: string
|-- image: string
|-- identifiers: array
|    |-- element: struct
|    |    |-- scheme: string
|    |    |-- identifier: string
|-- other_names: array
|    |-- element: struct
|    |    |-- lang: string
|    |    |-- note: string
|    |    |-- name: string
|-- sort_name: string
|-- images: array
|    |-- element: struct
|    |    |-- url: string
|-- given_name: string
|-- birth_date: string
|-- id: string
|-- contact_details: array
|    |-- element: struct
|    |    |-- type: string
|    |    |-- value: string
|-- death_date: string

Schema for the names DynamicFrame:
root
|-- family_name: string
|-- given_name: string

+-----------+----------+
|family_name|given_name|
+-----------+----------+
|    Collins|   Michael|
|   Huizenga|      Bill|
|    Clawson|    Curtis|
|    Solomon|    Gerald|
|     Rigell|    Edward|
|      Crapo|   Michael|
|      Hutto|      Earl|
|      Ertel|     Allen|
|     Minish|    Joseph|
|    Andrews|    Robert|
|     Walden|      Greg|
|      Kazen|   Abraham|
|     Turner|   Michael|
|      Kolbe|     James|
|  Lowenthal|      Alan|
|    Capuano|   Michael|
|   Schrader|      Kurt|
|     Nadler|   Jerrold|
|     Graves|       Tom|
|   McMillan|      John|
+-----------+----------+
only showing top 20 rows
```

simplify_ddb_json

`simplify_ddb_json(): DynamicFrame`

Simplifies nested columns in a `DynamicFrame` that are specifically in the DynamoDB JSON structure, and returns a new simplified `DynamicFrame`. If there’re multiple types or Map type in a List type, the elements in the List will not be simplified. Note that this is a specific type of transform that behaves differently from the regular `unnest` transform and requires the data to already be in the DynamoDB JSON structure. For more information, see DynamoDB JSON.

For example, the schema of a reading an export with the DynamoDB JSON structure might look like the following:

```
root
|-- Item: struct
|    |-- parentMap: struct
|    |    |-- M: struct
|    |    |    |-- childMap: struct
|    |    |    |    |-- M: struct
|    |    |    |    |    |-- appName: struct
|    |    |    |    |    |    |-- S: string
|    |    |    |    |    |-- packageName: struct
|    |    |    |    |    |    |-- S: string
|    |    |    |    |    |-- updatedAt: struct
|    |    |    |    |    |    |-- N: string
|    |-- strings: struct
|    |    |-- SS: array
|    |    |    |-- element: string
|    |-- numbers: struct
|    |    |-- NS: array
|    |    |    |-- element: string
|    |-- binaries: struct
|    |    |-- BS: array
|    |    |    |-- element: string
|    |-- isDDBJson: struct
|    |    |-- BOOL: boolean
|    |-- nullValue: struct
|    |    |-- NULL: boolean
```

The `simplify_ddb_json()` transform would convert this to:

```
root
|-- parentMap: struct
|    |-- childMap: struct
|    |    |-- appName: string
|    |    |-- packageName: string
|    |    |-- updatedAt: string
|-- strings: array
```

```
|    |-- element: string
|-- numbers: array
|    |-- element: string
|-- binaries: array
|    |-- element: string
|-- isDDBJson: boolean
|-- nullValue: null
```

Example: Use simplify_ddb_json to invoke a DynamoDB JSON simplify

This code example uses the `simplify_ddb_json` method to use the AWS Glue DynamoDB export connector, invoke a DynamoDB JSON simplify, and print the number of partitions.

Example code

```
from pyspark.context import SparkContext
from awsglue.context import GlueContext

sc = SparkContext()
glueContext = GlueContext(sc)

dynamicFrame = glueContext.create_dynamic_frame.from_options(
    connection_type = "dynamodb",
    connection_options = {
        'dynamodb.export': 'ddb',
        'dynamodb.tableArn': '<table arn>',
        'dynamodb.s3.bucket': '<bucket name>',
        'dynamodb.s3.prefix': '<bucket prefix>',
        'dynamodb.s3.bucketOwner': '<account_id of bucket>'
    }
)
simplified = dynamicFrame.simplify_ddb_json()
print(simplified.getNumPartitions())
```

spigot

`spigot(path, options={})`

Writes sample records to a specified destination to help you verify the transformations performed by your job.

- `path` – The path of the destination to write to (required).

- `options` – Key-value pairs that specify options (optional). The `"topk"` option specifies that the first k records should be written. The `"prob"` option specifies the probability (as a decimal) of choosing any given record. You can use it in selecting records to write.
- `transformation_ctx` – A unique string that is used to identify state information (optional).

Example: Use spigot to write sample fields from a `DynamicFrame` to Amazon S3

This code example uses the `spigot` method to write sample records to an Amazon S3 bucket after applying the `select_fields` transform.

Example dataset

> **Note**
>
> To access the dataset that is used in this example, see Code example: Joining and relationalizing data and follow the instructions in Step 1: Crawl the data in the Amazon S3 bucket.

The example uses a `DynamicFrame` called `persons` with the following schema:

```
root
|-- family_name: string
|-- name: string
|-- links: array
|    |-- element: struct
|    |    |-- note: string
|    |    |-- url: string
|-- gender: string
|-- image: string
|-- identifiers: array
|    |-- element: struct
|    |    |-- scheme: string
|    |    |-- identifier: string
|-- other_names: array
|    |-- element: struct
|    |    |-- lang: string
|    |    |-- note: string
|    |    |-- name: string
|-- sort_name: string
|-- images: array
```

```
|    |-- element: struct
|    |    |-- url: string
|-- given_name: string
|-- birth_date: string
|-- id: string
|-- contact_details: array
|    |-- element: struct
|    |    |-- type: string
|    |    |-- value: string
|-- death_date: string
```

Example code

```
# Example: Use spigot to write sample records
# to a destination during a transformation
# from pyspark.context import SparkContext.
# Replace DOC-EXAMPLE-BUCKET with your own location.

from pyspark.context import SparkContext
from awsglue.context import GlueContext

# Create GlueContext
sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)

# Load table data into a DynamicFrame
persons = glueContext.create_dynamic_frame.from_catalog(
    database="legislators", table_name="persons_json"
)

# Perform the select_fields on the DynamicFrame
persons = persons.select_fields(paths=["family_name", "given_name", "birth_date"])

# Use spigot to write a sample of the transformed data
# (the first 10 records)
spigot_output = persons.spigot(
    path="s3://DOC-EXAMPLE-BUCKET", options={"topk": 10}
)
# Example: Use spigot to write sample records
# to a destination during a transformation
# from pyspark.context import SparkContext.
# Replace DOC-EXAMPLE-BUCKET with your own location.
```

```
from pyspark.context import SparkContext
from awsglue.context import GlueContext

# Create GlueContext
sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)

# Load table data into a DynamicFrame
persons = glueContext.create_dynamic_frame.from_catalog(
    database="legislators", table_name="persons_json"
)

# Perform the select_fields on the DynamicFrame
persons = persons.select_fields(paths=["family_name", "given_name", "birth_date"])

# Use spigot to write a sample of the transformed data
# (the first 10 records)
spigot_output = persons.spigot(
    path="s3://DOC-EXAMPLE-BUCKET", options={"topk": 10}
)
```

Output

The following is an example of the data that `spigot` writes to Amazon S3. Because the example code specified `options={"topk": 10}`, the sample data contains the first 10 records.

```
{"family_name":"Collins","given_name":"Michael","birth_date":"1944-10-15"}
{"family_name":"Huizenga","given_name":"Bill","birth_date":"1969-01-31"}
{"family_name":"Clawson","given_name":"Curtis","birth_date":"1959-09-28"}
{"family_name":"Solomon","given_name":"Gerald","birth_date":"1930-08-14"}
{"family_name":"Rigell","given_name":"Edward","birth_date":"1960-05-28"}
{"family_name":"Crapo","given_name":"Michael","birth_date":"1951-05-20"}
{"family_name":"Hutto","given_name":"Earl","birth_date":"1926-05-12"}
{"family_name":"Ertel","given_name":"Allen","birth_date":"1937-11-07"}
{"family_name":"Minish","given_name":"Joseph","birth_date":"1916-09-01"}
{"family_name":"Andrews","given_name":"Robert","birth_date":"1957-08-04"}
```

split_fields

`split_fields(paths, name1, name2, transformation_ctx="", info="", stageThreshold=0, totalThreshold=0)`

Returns a new `DynamicFrameCollection` that contains two `DynamicFrames`. The first `DynamicFrame` contains all the nodes that have been split off, and the second contains the nodes that remain.

- `paths` – A list of strings, each of which is a full path to a node that you want to split into a new `DynamicFrame`.
- `name1` – A name string for the `DynamicFrame` that is split off.
- `name2` – A name string for the `DynamicFrame` that remains after the specified nodes have been split off.
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string to be associated with error reporting for this transformation (optional).
- `stageThreshold` – The number of errors encountered during this transformation at which the process should error out (optional). The default is zero, which indicates that the process should not error out.
- `totalThreshold` – The number of errors encountered up to and including this transformation at which the process should error out (optional). The default is zero, which indicates that the process should not error out.

Example: Use split_fields to split selected fields into a separate `DynamicFrame`

This code example uses the `split_fields` method to split a list of specified fields into a separate `DynamicFrame`.

Example dataset

The example uses a `DynamicFrame` called `l_root_contact_details` that is from a collection named `legislators_relationalized`.

`l_root_contact_details` has the following schema and entries.

```
root
|-- id: long
|-- index: int
|-- contact_details.val.type: string
|-- contact_details.val.value: string

+---+-----+------------------------+-------------------------+
| id|index|contact_details.val.type|contact_details.val.value|
+---+-----+------------------------+-------------------------+
|  1|    0|                   phone|             202-225-5265|
|  1|    1|                 twitter|              kathyhochul|
|  2|    0|                   phone|             202-225-3252|
|  2|    1|                 twitter|            repjackyrosen|
|  3|    0|                     fax|             202-225-1314|
|  3|    1|                   phone|             202-225-3772|
...
```

Example code

```
# Example: Use split_fields to split selected
# fields into a separate DynamicFrame

from pyspark.context import SparkContext
from awsglue.context import GlueContext

# Create GlueContext
sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)

# Load the input DynamicFrame and inspect its schema
frame_to_split = legislators_relationalized.select("l_root_contact_details")
print("Inspect the input DynamicFrame schema:")
frame_to_split.printSchema()

# Split id and index fields into a separate DynamicFrame
split_fields_collection = frame_to_split.split_fields(["id", "index"], "left", "right")

# Inspect the resulting DynamicFrames
print("Inspect the schemas of the DynamicFrames created with split_fields:")
split_fields_collection.select("left").printSchema()
split_fields_collection.select("right").printSchema()
```

Output

```
Inspect the input DynamicFrame's schema:
```

```
root
|-- id: long
|-- index: int
|-- contact_details.val.type: string
|-- contact_details.val.value: string

Inspect the schemas of the DynamicFrames created with split_fields:
root
|-- id: long
|-- index: int

root
|-- contact_details.val.type: string
|-- contact_details.val.value: string
```

split_rows

`split_rows(comparison_dict, name1, name2, transformation_ctx="", info="", stageThreshold=0, totalThreshold=0)`

Splits one or more rows in a `DynamicFrame` off into a new `DynamicFrame`.

The method returns a new `DynamicFrameCollection` that contains two `DynamicFrames`. The first `DynamicFrame` contains all the rows that have been split off, and the second contains the rows that remain.

- `comparison_dict` – A dictionary where the key is a path to a column, and the value is another dictionary for mapping comparators to values that the column values are compared to. For example, `{"age": {">": 10, "<": 20}}` splits off all rows whose value in the age column is greater than 10 and less than 20.
- `name1` – A name string for the `DynamicFrame` that is split off.
- `name2` – A name string for the `DynamicFrame` that remains after the specified nodes have been split off.
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string to be associated with error reporting for this transformation (optional).
- `stageThreshold` – The number of errors encountered during this transformation at which the process should error out (optional). The default is zero, which indicates that the process should not error out.

- `totalThreshold` – The number of errors encountered up to and including this transformation at which the process should error out (optional). The default is zero, which indicates that the process should not error out.

Example: Use split_rows to split rows in a `DynamicFrame`

This code example uses the `split_rows` method to split rows in a `DynamicFrame` based on the `id` field value.

Example dataset

The example uses a `DynamicFrame` called `l_root_contact_details` that is selected from a collection named `legislators_relationalized`.

`l_root_contact_details` has the following schema and entries.

```
root
|-- id: long
|-- index: int
|-- contact_details.val.type: string
|-- contact_details.val.value: string

+---+-----+------------------------+-------------------------+
| id|index|contact_details.val.type|contact_details.val.value|
+---+-----+------------------------+-------------------------+
|  1|    0|                   phone|             202-225-5265|
|  1|    1|                 twitter|              kathyhochul|
|  2|    0|                   phone|             202-225-3252|
|  2|    1|                 twitter|            repjackyrosen|
|  3|    0|                     fax|             202-225-1314|
|  3|    1|                   phone|             202-225-3772|
|  3|    2|                 twitter|          MikeRossUpdates|
|  4|    0|                     fax|             202-225-1314|
|  4|    1|                   phone|             202-225-3772|
|  4|    2|                 twitter|          MikeRossUpdates|
|  5|    0|                     fax|             202-225-1314|
|  5|    1|                   phone|             202-225-3772|
|  5|    2|                 twitter|          MikeRossUpdates|
|  6|    0|                     fax|             202-225-1314|
|  6|    1|                   phone|             202-225-3772|
|  6|    2|                 twitter|          MikeRossUpdates|
|  7|    0|                     fax|             202-225-1314|
|  7|    1|                   phone|             202-225-3772|
|  7|    2|                 twitter|          MikeRossUpdates|
|  8|    0|                     fax|             202-225-1314|
+---+-----+------------------------+-------------------------+
```

Example code

```
# Example: Use split_rows to split up
# rows in a DynamicFrame based on value

from pyspark.context import SparkContext
from awsglue.context import GlueContext

# Create GlueContext
sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)

# Retrieve the DynamicFrame to split
frame_to_split = legislators_relationalized.select("l_root_contact_details")

# Split up rows by ID
split_rows_collection = frame_to_split.split_rows({"id": {">": 10}}, "high", "low")

# Inspect the resulting DynamicFrames
print("Inspect the DynamicFrame that contains IDs < 10")
split_rows_collection.select("low").toDF().show()
print("Inspect the DynamicFrame that contains IDs > 10")
split_rows_collection.select("high").toDF().show()
```

Output

```
Inspect the DynamicFrame that contains IDs < 10
+---+-----+------------------------+-------------------------+
| id|index|contact_details.val.type|contact_details.val.value|
+---+-----+------------------------+-------------------------+
|  1|    0|                   phone|             202-225-5265|
|  1|    1|                 twitter|              kathyhochul|
|  2|    0|                   phone|             202-225-3252|
|  2|    1|                 twitter|            repjackyrosen|
|  3|    0|                     fax|             202-225-1314|
|  3|    1|                   phone|             202-225-3772|
|  3|    2|                 twitter|          MikeRossUpdates|
|  4|    0|                     fax|             202-225-1314|
|  4|    1|                   phone|             202-225-3772|
|  4|    2|                 twitter|          MikeRossUpdates|
|  5|    0|                     fax|             202-225-1314|
|  5|    1|                   phone|             202-225-3772|
|  5|    2|                 twitter|          MikeRossUpdates|
|  6|    0|                     fax|             202-225-1314|
|  6|    1|                   phone|             202-225-3772|
|  6|    2|                 twitter|          MikeRossUpdates|
|  7|    0|                     fax|             202-225-1314|
|  7|    1|                   phone|             202-225-3772|
|  7|    2|                 twitter|          MikeRossUpdates|
|  8|    0|                     fax|             202-225-1314|
+---+-----+------------------------+-------------------------+
only showing top 20 rows

Inspect the DynamicFrame that contains IDs > 10
+---+-----+------------------------+-------------------------+
| id|index|contact_details.val.type|contact_details.val.value|
+---+-----+------------------------+-------------------------+
| 11|    0|                   phone|             202-225-5476|
| 11|    1|                 twitter|            RepDavidYoung|
| 12|    0|                   phone|             202-225-4035|
| 12|    1|                 twitter|           RepStephMurphy|
| 13|    0|                     fax|             202-226-0774|
| 13|    1|                   phone|             202-225-6335|
| 14|    0|                     fax|             202-226-0774|
| 14|    1|                   phone|             202-225-6335|
| 15|    0|                     fax|             202-226-0774|
| 15|    1|                   phone|             202-225-6335|
| 16|    0|                     fax|             202-226-0774|
| 16|    1|                   phone|             202-225-6335|
| 17|    0|                     fax|             202-226-0774|
| 17|    1|                   phone|             202-225-6335|
| 18|    0|                     fax|             202-226-0774|
| 18|    1|                   phone|             202-225-6335|
| 19|    0|                     fax|             202-226-0774|
| 19|    1|                   phone|             202-225-6335|
| 20|    0|                     fax|             202-226-0774|
| 20|    1|                   phone|             202-225-6335|
+---+-----+------------------------+-------------------------+
only showing top 20 rows
```

unbox

`unbox(path, format, transformation_ctx="", info="", stageThreshold=0, totalThreshold=0, **options)`

Unboxes (reformats) a string field in a `DynamicFrame` and returns a new `DynamicFrame` that contains the unboxed `DynamicRecords`.

A `DynamicRecord` represents a logical record in a `DynamicFrame`. It's similar to a row in an Apache Spark `DataFrame`, except that it is self-describing and can be used for data that doesn't conform to a fixed schema.

- `path` – A full path to the string node you want to unbox.
- `format` – A format specification (optional). You use this for an Amazon S3 or AWS Glue connection that supports multiple formats. For the formats that are supported, see Data format options for inputs and outputs in AWS Glue for Spark.
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string to be associated with error reporting for this transformation (optional).
- `stageThreshold` – The number of errors encountered during this transformation at which the process should error out (optional). The default is zero, which indicates that the process should not error out.
- `totalThreshold` – The number of errors encountered up to and including this transformation at which the process should error out (optional). The default is zero, which indicates that the process should not error out.
- `options` – One or more of the following:
  - `separator` – A string that contains the separator character.
  - `escaper` – A string that contains the escape character.
  - `skipFirst` – A Boolean value that indicates whether to skip the first instance.
  - `withSchema` – A string containing a JSON representation of the node's schema. The format of a schema's JSON representation is defined by the output of `StructType.json()`.
  - `withHeader` – A Boolean value that indicates whether a header is included.

Example: Use unbox to unbox a string field into a struct

This code example uses the `unbox` method to *unbox*, or reformat, a string field in a `DynamicFrame` into a field of type struct.

Example dataset

The example uses a `DynamicFrame` called `mapped_with_string` with the following schema and entries.

Notice the field named `AddressString`. This is the field that the example unboxes into a struct.

```
root
|-- Average Total Payments: string
|-- AddressString: string
|-- Average Covered Charges: string
|-- DRG Definition: string
|-- Average Medicare Payments: string
|-- Hospital Referral Region Description: string
|-- Address: struct
|    |-- Zip.Code: string
|    |-- City: string
|    |-- Array: array
|    |    |-- element: string
|    |-- State: string
|    |-- Street: string
|-- Provider Id: string
|-- Total Discharges: string
|-- Provider Name: string

+----------------------+--------------------+-----------------------
+--------------------+-------------------------+------------------------------------
+--------------------+-----------+----------------+--------------------+
|Average Total Payments|       AddressString|Average Covered Charges|      DRG
 Definition|Average Medicare Payments|Hospital Referral Region Description|
 Address|Provider Id|Total Discharges|       Provider Name|
+----------------------+--------------------+-----------------------
+--------------------+-------------------------+------------------------------------
+--------------------+-----------+----------------+--------------------+
|              $5777.24|{"Street": "1108 ...|              $32963.07|039 -
 EXTRACRANIA...|                 $4763.73|                         AL - Dothan|[36301,
 DOTHAN, [...|      10001|              91|SOUTHEAST ALABAMA...|
|              $5787.57|{"Street": "2505 ...|              $15131.85|039 -
 EXTRACRANIA...|                 $4976.71|                     AL - Birmingham|[35957,
 BOAZ, [25...|      10005|              14|MARSHALL MEDICAL ...|
|              $5434.95|{"Street": "205 M...|              $37560.37|039 -
 EXTRACRANIA...|                 $4453.79|                     AL - Birmingham|[35631,
 FLORENCE,...|      10006|              24|ELIZA COFFEE MEMO...|
|              $5417.56|{"Street": "50 ME...|              $13998.28|039 -
 EXTRACRANIA...|                 $4129.16|                     AL - Birmingham|[35235,
 BIRMINGHA...|      10011|              25|   ST VINCENT'S EAST|
...
```

Example code

```
# Example: Use unbox to unbox a string field
# into a struct in a DynamicFrame

from pyspark.context import SparkContext
from awsglue.context import GlueContext

# Create GlueContext
sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)

unboxed = mapped_with_string.unbox("AddressString", "json")
unboxed.printSchema()
unboxed.toDF().show()
```

Output

```
root
|-- Average Total Payments: string
|-- AddressString: struct
|    |-- Street: string
|    |-- City: string
|    |-- State: string
|    |-- Zip.Code: string
|    |-- Array: array
|    |    |-- element: string
|-- Average Covered Charges: string
|-- DRG Definition: string
|-- Average Medicare Payments: string
|-- Hospital Referral Region Description: string
|-- Address: struct
|    |-- Zip.Code: string
|    |-- City: string
|    |-- Array: array
|    |    |-- element: string
|    |-- State: string
|    |-- Street: string
|-- Provider Id: string
|-- Total Discharges: string
|-- Provider Name: string

+----------------------+--------------------+-----------------------
+--------------------+-------------------------+------------------------------------
+--------------------+-----------+----------------+--------------------+
|Average Total Payments|       AddressString|Average Covered Charges|      DRG
 Definition|Average Medicare Payments|Hospital Referral Region Description|
 Address|Provider Id|Total Discharges|       Provider Name|
+----------------------+--------------------+-----------------------
+--------------------+-------------------------+------------------------------------
+--------------------+-----------+----------------+--------------------+
|              $5777.24|[1108 ROSS CLARK ...|              $32963.07|039 -
 EXTRACRANIA...|                 $4763.73|                         AL - Dothan|[36301,
 DOTHAN, [...|      10001|              91|SOUTHEAST ALABAMA...|
|              $5787.57|[2505 U S HIGHWAY...|              $15131.85|039 -
 EXTRACRANIA...|                 $4976.71|                     AL - Birmingham|[35957,
 BOAZ, [25...|      10005|              14|MARSHALL MEDICAL ...|
|              $5434.95|[205 MARENGO STRE...|              $37560.37|039 -
 EXTRACRANIA...|                 $4453.79|                     AL - Birmingham|[35631,
 FLORENCE,...|      10006|              24|ELIZA COFFEE MEMO...|
|              $5417.56|[50 MEDICAL PARK ...|              $13998.28|039 -
 EXTRACRANIA...|                 $4129.16|                     AL - Birmingham|[35235,
 BIRMINGHA...|      10011|              25|   ST VINCENT'S EAST|
|              $5658.33|[1000 FIRST STREE...|              $31633.27|039 -
 EXTRACRANIA...|                 $4851.44|                     AL - Birmingham|[35007,
 ALABASTER...|      10016|              18|SHELBY BAPTIST ME...|
|              $6653.80|[2105 EAST SOUTH ...|              $16920.79|039 -
 EXTRACRANIA...|                 $5374.14|                     AL - Montgomery|[36116,
 MONTGOMER...|      10023|              67|BAPTIST MEDICAL C...|
|              $5834.74|[2000 PEPPERELL P...|              $11977.13|039 -
 EXTRACRANIA...|                 $4761.41|                     AL - Birmingham|[36801,
 OPELIKA, ...|      10029|              51|EAST ALABAMA MEDI...|
|              $8031.12|[619 SOUTH 19TH S...|              $35841.09|039 -
 EXTRACRANIA...|                 $5858.50|                     AL - Birmingham|[35233,
 BIRMINGHA...|      10033|              32|UNIVERSITY OF ALA...|
|              $6113.38|[101 SIVLEY RD, H...|              $28523.39|039 -
 EXTRACRANIA...|                 $5228.40|                     AL - Huntsville|[35801,
 HUNTSVILL...|      10039|             135| HUNTSVILLE HOSPITAL|
|              $5541.05|[1007 GOODYEAR AV...|              $75233.38|039 -
 EXTRACRANIA...|                 $4386.94|                     AL - Birmingham|[35903,
 GADSDEN, ...|      10040|              34|GADSDEN REGIONAL ...|
|              $5461.57|[600 SOUTH THIRD ...|              $67327.92|039 -
 EXTRACRANIA...|                $4493.57|                    AL - Birmingham|[35901,
 GADSDEN, ...|      10046|              14|RIVERVIEW REGIONA...|
|              $5356.28|[4370 WEST MAIN S...|              $39607.28|039 -
 EXTRACRANIA...|                $4408.20|                        AL - Dothan|[36305,
 DOTHAN, [...|      10055|              45|    FLOWERS HOSPITAL|
|              $5374.65|[810 ST VINCENT'S...|              $22862.23|039 -
 EXTRACRANIA...|                $4186.02|                    AL - Birmingham|[35205,
 BIRMINGHA...|      10056|              43|ST VINCENT'S BIRM...|
|              $5366.23|[400 EAST 10TH ST...|              $31110.85|039 -
 EXTRACRANIA...|                $4376.23|                    AL - Birmingham|[36207,
 ANNISTON,...|      10078|              21|NORTHEAST ALABAMA...|
|              $5282.93|[1613 NORTH MCKEN...|              $25411.33|039 -
 EXTRACRANIA...|                $4383.73|                        AL - Mobile|[36535,
 FOLEY, [1...|      10083|              15|SOUTH BALDWIN REG...|
|              $5676.55|[1201 7TH STREET ...|               $9234.51|039 -
 EXTRACRANIA...|                $4509.11|                    AL - Huntsville|[35609,
 DECATUR, ...|      10085|              27|DECATUR GENERAL H...|
|              $5930.11|[6801 AIRPORT BOU...|              $15895.85|039 -
 EXTRACRANIA...|                $3972.85|                        AL - Mobile|[36608,
 MOBILE, [...|      10090|              27| PROVIDENCE HOSPITAL|
|              $6192.54|[809 UNIVERSITY B...|              $19721.16|039 -
 EXTRACRANIA...|                $5179.38|                    AL - Tuscaloosa|[35401,
 TUSCALOOS...|      10092|              31|D C H REGIONAL ME...|
|              $4968.00|[750 MORPHY AVENU...|              $10710.88|039 -
 EXTRACRANIA...|                $3898.88|                        AL - Mobile|[36532,
 FAIRHOPE,...|      10100|              18|     THOMAS HOSPITAL|
|              $5996.00|[701 PRINCETON AV...|              $51343.75|039 -
 EXTRACRANIA...|                $4962.45|                    AL - Birmingham|[35211,
 BIRMINGHA...|      10103|              33|BAPTIST MEDICAL C...|
+----------------------+--------------------+-----------------------
+--------------------+-------------------------+------------------------------------
+--------------------+-----------+----------------+--------------------+
only showing top 20 rows
```

union

`union(frame1, frame2, transformation_ctx = "", info = "", stageThreshold = 0, totalThreshold = 0)`

Union two DynamicFrames. Returns DynamicFrame containing all records from both input DynamicFrames. This transform may return different results from the union of two DataFrames with equivalent data. If you need the Spark DataFrame union behavior, consider using `toDF`.

- `frame1` – First DynamicFrame to union.
- `frame2` – Second DynamicFrame to union.
- `transformation_ctx` – (optional) A unique string that is used to identify stats / state information
- `info` – (optional) Any string to be associated with errors in the transformation
- `stageThreshold` – (optional) Max number of errors in the transformation until processing will error out
- `totalThreshold` – (optional) Max number of errors total until processing will error out.

unnest

`unnest(transformation_ctx="", info="", stageThreshold=0, totalThreshold=0)`

Unnests nested objects in a `DynamicFrame`, which makes them top-level objects, and returns a new unnested `DynamicFrame`.

- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string to be associated with error reporting for this transformation (optional).
- `stageThreshold` – The number of errors encountered during this transformation at which the process should error out (optional). The default is zero, which indicates that the process should not error out.
- `totalThreshold` – The number of errors encountered up to and including this transformation at which the process should error out (optional). The default is zero, which indicates that the process should not error out.

Example: Use unnest to turn nested fields into top-level fields

This code example uses the `unnest` method to flatten all of the nested fields in a `DynamicFrame` into top-level fields.

Example dataset

The example uses a `DynamicFrame` called `mapped_medicare` with the following schema. Notice that the `Address` field is the only field that contains nested data.

```
root
|-- Average Total Payments: string
|-- Average Covered Charges: string
|-- DRG Definition: string
|-- Average Medicare Payments: string
|-- Hospital Referral Region Description: string
|-- Address: struct
|    |-- Zip.Code: string
|    |-- City: string
|    |-- Array: array
|    |    |-- element: string
|    |-- State: string
|    |-- Street: string
|-- Provider Id: string
|-- Total Discharges: string
|-- Provider Name: string
```

Example code

```
# Example: Use unnest to unnest nested
# objects in a DynamicFrame

from pyspark.context import SparkContext
from awsglue.context import GlueContext

# Create GlueContext
sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)

# Unnest all nested fields
unnested = mapped_medicare.unnest()
unnested.printSchema()
```

Output

```
root
|-- Average Total Payments: string
|-- Average Covered Charges: string
|-- DRG Definition: string
|-- Average Medicare Payments: string
|-- Hospital Referral Region Description: string
|-- Address.Zip.Code: string
|-- Address.City: string
|-- Address.Array: array
|    |-- element: string
|-- Address.State: string
|-- Address.Street: string
|-- Provider Id: string
|-- Total Discharges: string
|-- Provider Name: string
```

unnest_ddb_json

Unnests nested columns in a `DynamicFrame` that are specifically in the DynamoDB JSON structure, and returns a new unnested `DynamicFrame`. Columns that are of an array of struct types will not be unnested. Note that this is a specific type of unnesting transform that behaves differently from the regular `unnest` transform and requires the data to already be in the DynamoDB JSON structure. For more information, see DynamoDB JSON.

`unnest_ddb_json(transformation_ctx="", info="", stageThreshold=0, totalThreshold=0)`

- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string to be associated with error reporting for this transformation (optional).
- `stageThreshold` – The number of errors encountered during this transformation at which the process should error out (optional: zero by default, indicating that the process should not error out).
- `totalThreshold` – The number of errors encountered up to and including this transformation at which the process should error out (optional: zero by default, indicating that the process should not error out).

For example, the schema of a reading an export with the DynamoDB JSON structure might look like the following:

```
root
|-- Item: struct
|    |-- ColA: struct
|    |    |-- S: string
|    |-- ColB: struct
|    |    |-- S: string
|    |-- ColC: struct
|    |    |-- N: string
|    |-- ColD: struct
|    |    |-- L: array
|    |    |    |-- element: null
```

The `unnest_ddb_json()` transform would convert this to:

```
root
|-- ColA: string
|-- ColB: string
|-- ColC: string
|-- ColD: array
|    |-- element: null
```

The following code example shows how to use the AWS Glue DynamoDB export connector, invoke a DynamoDB JSON unnest, and print the number of partitions:

```
import sys
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job
from awsglue.utils import getResolvedOptions

args = getResolvedOptions(sys.argv, ["JOB_NAME"])
glue_context= GlueContext(SparkContext.getOrCreate())
job = Job(glue_context)
job.init(args["JOB_NAME"], args)

dynamicFrame = glue_context.create_dynamic_frame.from_options(
    connection_type="dynamodb",
    connection_options={
        "dynamodb.export": "ddb",
        "dynamodb.tableArn": "<test_source>",
        "dynamodb.s3.bucket": "<bucket name>",
        "dynamodb.s3.prefix": "<bucket prefix>",
        "dynamodb.s3.bucketOwner": "<account_id>",
    }
)
unnested = dynamicFrame.unnest_ddb_json()
print(unnested.getNumPartitions())

job.commit()
```

write

`write(connection_type, connection_options, format, format_options, accumulator_size)`

Gets a DataSink(object) of the specified connection type from the GlueContext class of this `DynamicFrame`, and uses it to format and write the contents of this `DynamicFrame`. Returns the new `DynamicFrame` formatted and written as specified.

- `connection_type` – The connection type to use. Valid values include `s3`, `mysql`, `postgresql`, `redshift`, `sqlserver`, and `oracle`.
- `connection_options` – The connection option to use (optional). For a `connection_type` of `s3`, an Amazon S3 path is defined.

  ```
  connection_options = {"path": "s3://aws-glue-target/temp"}
  ```

  For JDBC connections, several properties must be defined. Note that the database name must be part of the URL. It can optionally be included in the connection options.

  > **Warning**
  >
  > Storing passwords in your script is not recommended. Consider using `boto3` to retrieve them from AWS Secrets Manager or the AWS Glue Data Catalog.

  ```
  connection_options = {"url": "jdbc-url/database", "user": "username",
   "password": passwordVariable,"dbtable": "table-name", "redshiftTmpDir": "s3-tempdir-
  path"}
  ```

- `format` – A format specification (optional). This is used for an Amazon Simple Storage Service (Amazon S3) or an AWS Glue connection that supports multiple formats. See Data format options for inputs and outputs in AWS Glue for Spark for the formats that are supported.
- `format_options` – Format options for the specified format. See Data format options for inputs and outputs in AWS Glue for Spark for the formats that are supported.
- `accumulator_size` – The accumulable size to use, in bytes (optional).

— errors —

- assertErrorThreshold
- errorsAsDynamicFrame
- errorsCount
- stageErrorsCount

assertErrorThreshold

`assertErrorThreshold( )` – An assert for errors in the transformations that created this `DynamicFrame`. Returns an `Exception` from the underlying `DataFrame`.

errorsAsDynamicFrame

`errorsAsDynamicFrame( )` – Returns a `DynamicFrame` that has error records nested inside.

Example: Use errorsAsDynamicFrame to view error records

The following code example shows how to use the `errorsAsDynamicFrame` method to view an error record for a `DynamicFrame`.

Example dataset

The example uses the following dataset that you can upload to Amazon S3 as JSON. Notice that the second record is malformed. Malformed data typically breaks file parsing when you use SparkSQL. However, `DynamicFrame` recognizes malformation issues and turns malformed lines into error records that you can handle individually.

```
{"id": 1, "name": "george", "surname": "washington", "height": 178}
{"id": 2, "name": "benjamin", "surname": "franklin",
{"id": 3, "name": "alexander", "surname": "hamilton", "height": 171}
{"id": 4, "name": "john", "surname": "jay", "height": 190}
```

Example code

```
# Example: Use errorsAsDynamicFrame to view error records.
# Replace s3://DOC-EXAMPLE-S3-BUCKET/error_data.json with your location.

from pyspark.context import SparkContext
from awsglue.context import GlueContext

# Create GlueContext
sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)

# Create errors DynamicFrame, view schema
errors = glueContext.create_dynamic_frame.from_options(
    "s3", {"paths": ["s3://DOC-EXAMPLE-S3-BUCKET/error_data.json"]}, "json"
)
print("Schema of errors DynamicFrame:")
errors.printSchema()

# Show that errors only contains valid entries from the dataset
print("errors contains only valid records from the input dataset (2 of 4 records)")
errors.toDF().show()

# View errors
print("Errors count:", str(errors.errorsCount()))
print("Errors:")
errors.errorsAsDynamicFrame().toDF().show()

# View error fields and error data
error_record = errors.errorsAsDynamicFrame().toDF().head()

error_fields = error_record["error"]
print("Error fields: ")
print(error_fields.asDict().keys())

print("\nError record data:")
for key in error_fields.asDict().keys():
    print("\n", key, ": ", str(error_fields[key]))
```

Output

```
Schema of errors DynamicFrame:
root
```

```
|-- id: int
|-- name: string
|-- surname: string
|-- height: int

errors contains only valid records from the input dataset (2 of 4 records)
+---+------+----------+------+
| id|  name|   surname|height|
+---+------+----------+------+
|  1|george|washington|   178|
|  4|  john|       jay|   190|
+---+------+----------+------+

Errors count: 1
Errors:
+--------------------+
|               error|
+--------------------+
|[[  File "/tmp/20...|
+--------------------+

Error fields:
dict_keys(['callsite', 'msg', 'stackTrace', 'input', 'bytesread', 'source',
 'dynamicRecord'])

Error record data:

 callsite :  Row(site='  File "/tmp/2060612586885849088", line 549, in <module>\n
 sys.exit(main())\n  File "/tmp/2060612586885849088", line 523, in main\n    response
 = handler(content)\n  File "/tmp/2060612586885849088", line 197, in execute_request
\n    result = node.execute()\n  File "/tmp/2060612586885849088", line 103, in
 execute\n    exec(code, global_dict)\n  File "<stdin>", line 10, in <module>\n
 File "/opt/amazon/lib/python3.6/site-packages/awsglue/dynamicframe.py", line 625, in
 from_options\n    format_options, transformation_ctx, push_down_predicate, **kwargs)\n
  File "/opt/amazon/lib/python3.6/site-packages/awsglue/context.py", line 233, in
 create_dynamic_frame_from_options\n    source.setFormat(format, **format_options)\n',
 info='')

 msg :  error in jackson reader

 stackTrace :  com.fasterxml.jackson.core.JsonParseException: Unexpected character
 ('{' (code 123)): was expecting either valid name character (for unquoted name) or
 double-quote (for quoted) to start field name
 at [Source: com.amazonaws.services.glue.readers.BufferedStream@73492578; line: 3,
 column: 2]
 at com.fasterxml.jackson.core.JsonParser._constructError(JsonParser.java:1581)
 at
 com.fasterxml.jackson.core.base.ParserMinimalBase._reportError(ParserMinimalBase.java:533)
 at
 com.fasterxml.jackson.core.base.ParserMinimalBase._reportUnexpectedChar(ParserMinimalBase.java
 at
 com.fasterxml.jackson.core.json.UTF8StreamJsonParser._handleOddName(UTF8StreamJsonParser.java:
 at
 com.fasterxml.jackson.core.json.UTF8StreamJsonParser._parseName(UTF8StreamJsonParser.java:1650
 at
 com.fasterxml.jackson.core.json.UTF8StreamJsonParser.nextToken(UTF8StreamJsonParser.java:740)
 at com.amazonaws.services.glue.readers.JacksonReader$$anonfun$hasNextGoodToken
$1.apply(JacksonReader.scala:57)
 at com.amazonaws.services.glue.readers.JacksonReader$$anonfun$hasNextGoodToken
$1.apply(JacksonReader.scala:57)
 at scala.collection.Iterator$$anon$9.next(Iterator.scala:162)
 at scala.collection.Iterator$$anon$16.hasNext(Iterator.scala:599)
 at scala.collection.Iterator$$anon$16.hasNext(Iterator.scala:598)
 at scala.collection.Iterator$class.foreach(Iterator.scala:891)
 at scala.collection.AbstractIterator.foreach(Iterator.scala:1334)
 at com.amazonaws.services.glue.readers.JacksonReader$$anonfun
$1.apply(JacksonReader.scala:120)
 at com.amazonaws.services.glue.readers.JacksonReader$$anonfun
$1.apply(JacksonReader.scala:116)
 at
 com.amazonaws.services.glue.DynamicRecordBuilder.handleErr(DynamicRecordBuilder.scala:209)
 at
 com.amazonaws.services.glue.DynamicRecordBuilder.handleErrorWithException(DynamicRecordBuilder
 at
 com.amazonaws.services.glue.readers.JacksonReader.nextFailSafe(JacksonReader.scala:116)
 at com.amazonaws.services.glue.readers.JacksonReader.next(JacksonReader.scala:109)
 at com.amazonaws.services.glue.readers.JSONReader.next(JSONReader.scala:247)
 at
 com.amazonaws.services.glue.hadoop.TapeHadoopRecordReaderSplittable.nextKeyValue(TapeHadoopRec
 at org.apache.spark.rdd.NewHadoopRDD$$anon$1.hasNext(NewHadoopRDD.scala:230)
 at org.apache.spark.InterruptibleIterator.hasNext(InterruptibleIterator.scala:37)
 at scala.collection.Iterator$$anon$11.hasNext(Iterator.scala:409)
 at scala.collection.Iterator$$anon$11.hasNext(Iterator.scala:409)
 at scala.collection.Iterator$$anon$13.hasNext(Iterator.scala:462)
 at scala.collection.Iterator$$anon$11.hasNext(Iterator.scala:409)
 at scala.collection.Iterator$$anon$11.hasNext(Iterator.scala:409)
 at scala.collection.Iterator$$anon$13.hasNext(Iterator.scala:462)
 at scala.collection.Iterator$$anon$11.hasNext(Iterator.scala:409)
 at scala.collection.Iterator$$anon$11.hasNext(Iterator.scala:409)
 at org.apache.spark.sql.execution.SparkPlan$$anonfun$2.apply(SparkPlan.scala:255)
 at org.apache.spark.sql.execution.SparkPlan$$anonfun$2.apply(SparkPlan.scala:247)
 at org.apache.spark.rdd.RDD$$anonfun$mapPartitionsInternal$1$$anonfun$apply
$24.apply(RDD.scala:836)
 at org.apache.spark.rdd.RDD$$anonfun$mapPartitionsInternal$1$$anonfun$apply
$24.apply(RDD.scala:836)
 at org.apache.spark.rdd.MapPartitionsRDD.compute(MapPartitionsRDD.scala:52)
 at org.apache.spark.rdd.RDD.computeOrReadCheckpoint(RDD.scala:324)
 at org.apache.spark.rdd.RDD.iterator(RDD.scala:288)
 at org.apache.spark.rdd.MapPartitionsRDD.compute(MapPartitionsRDD.scala:52)
 at org.apache.spark.rdd.RDD.computeOrReadCheckpoint(RDD.scala:324)
 at org.apache.spark.rdd.RDD.iterator(RDD.scala:288)
 at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:90)
 at org.apache.spark.scheduler.Task.run(Task.scala:121)
 at org.apache.spark.executor.Executor$TaskRunner$$anonfun$10.apply(Executor.scala:408)
 at org.apache.spark.util.Utils$.tryWithSafeFinally(Utils.scala:1360)
 at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:414)
 at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
 at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
 at java.lang.Thread.run(Thread.java:750)


 input :

 bytesread :  252

 source :

 dynamicRecord :  Row(id=2, name='benjamin', surname='franklin')
```

errorsCount

`errorsCount( )` – Returns the total number of errors in a `DynamicFrame`.

stageErrorsCount

`stageErrorsCount` – Returns the number of errors that occurred in the process of generating this `DynamicFrame`.

DynamicFrameCollection class

A `DynamicFrameCollection` is a dictionary of DynamicFrame class objects, in which the keys are the names of the `DynamicFrames` and the values are the `DynamicFrame` objects.

__init__

`__init__(dynamic_frames, glue_ctx)`

- `dynamic_frames` – A dictionary of DynamicFrame class objects.
- `glue_ctx` – A GlueContext class object.

Keys

`keys( )` – Returns a list of the keys in this collection, which generally consists of the names of the corresponding `DynamicFrame` values.

Values

`values(key)` – Returns a list of the `DynamicFrame` values in this collection.

Select

`select(key)`

Returns the `DynamicFrame` that corresponds to the specfied key (which is generally the name of the `DynamicFrame`).

- `key` – A key in the `DynamicFrameCollection`, which usually represents the name of a `DynamicFrame`.

Map

`map(callable, transformation_ctx="")`

Uses a passed-in function to create and return a new `DynamicFrameCollection` based on the `DynamicFrames` in this collection.

- `callable` – A function that takes a `DynamicFrame` and the specified transformation context as parameters and returns a `DynamicFrame`.
- `transformation_ctx` – A transformation context to be used by the callable (optional).

Flatmap

`flatmap(f, transformation_ctx="")`

Uses a passed-in function to create and return a new `DynamicFrameCollection` based on the `DynamicFrames` in this collection.

- `f` – A function that takes a `DynamicFrame` as a parameter and returns a `DynamicFrame` or `DynamicFrameCollection`.
- `transformation_ctx` – A transformation context to be used by the function (optional).

DynamicFrameWriter class

methods

- __init__
- from_options
- from_catalog
- from_jdbc_conf

__init__

`__init__(glue_context)`

- `glue_context` – The GlueContext class to use.

from_options

`from_options(frame, connection_type, connection_options={}, format=None, format_options={}, transformation_ctx="")`

Writes a `DynamicFrame` using the specified connection and format.

- `frame` – The `DynamicFrame` to write.
- `connection_type` – The connection type. Valid values include `s3`, `mysql`, `postgresql`, `redshift`, `sqlserver`, and `oracle`.
- `connection_options` – Connection options, such as path and database table (optional). For a `connection_type` of `s3`, an Amazon S3 path is defined.

```
connection_options = {"path": "s3://aws-glue-target/temp"}
```

For JDBC connections, several properties must be defined. Note that the database name must be part of the URL. It can optionally be included in the connection options.

> **Warning**
>
> Storing passwords in your script is not recommended. Consider using `boto3` to retrieve them from AWS Secrets Manager or the AWS Glue Data Catalog.

```
connection_options = {"url": "jdbc-url/database", "user": "username",
 "password": passwordVariable,"dbtable": "table-name", "redshiftTmpDir": "s3-tempdir-
path"}
```

The `dbtable` property is the name of the JDBC table. For JDBC data stores that support schemas within a database, specify `schema.table-name`. If a schema is not provided, then the default "public" schema is used.

For more information, see Connection types and options for ETL in AWS Glue for Spark.

- `format` – A format specification (optional). This is used for an Amazon Simple Storage Service (Amazon S3) or an AWS Glue connection that supports multiple formats. See Data format options for inputs and outputs in AWS Glue for Spark for the formats that are supported.
- `format_options` – Format options for the specified format. See Data format options for inputs and outputs in AWS Glue for Spark for the formats that are supported.
- `transformation_ctx` – A transformation context to use (optional).

from_catalog

`from_catalog(frame, name_space, table_name, redshift_tmp_dir="", transformation_ctx="")`

Writes a `DynamicFrame` using the specified catalog database and table name.

- `frame` – The `DynamicFrame` to write.
- `name_space` – The database to use.

- `table_name` – The `table_name` to use.
- `redshift_tmp_dir` – An Amazon Redshift temporary directory to use (optional).
- `transformation_ctx` – A transformation context to use (optional).
- `additional_options` – Additional options provided to AWS Glue.

  To write to Lake Formation governed tables, you can use these additional options:

  - `transactionId` – (String) The transaction ID at which to do the write to the Governed table. This transaction can not be already committed or aborted, or the write will fail.
  - `callDeleteObjectsOnCancel` – (Boolean, optional) If set to `true` (default), AWS Glue automatically calls the `DeleteObjectsOnCancel` API after the object is written to Amazon S3. For more information, see DeleteObjectsOnCancel in the *AWS Lake Formation Developer Guide*.

  Example Example: Writing to a governed table in Lake Formation

```
txId = glueContext.start_transaction(read_only=False)
glueContext.write_dynamic_frame.from_catalog(
    frame=dyf,
    database = db,
    table_name = tbl,
    transformation_ctx = "datasource0",
    additional_options={"transactionId":txId})
...
glueContext.commit_transaction(txId)
```

from_jdbc_conf

`from_jdbc_conf(frame, catalog_connection, connection_options={}, redshift_tmp_dir = "", transformation_ctx="")`

Writes a `DynamicFrame` using the specified JDBC connection information.

- `frame` – The `DynamicFrame` to write.
- `catalog_connection` – A catalog connection to use.
- `connection_options` – Connection options, such as path and database table (optional).
- `redshift_tmp_dir` – An Amazon Redshift temporary directory to use (optional).
- `transformation_ctx` – A transformation context to use (optional).

Example for write_dynamic_frame

This example writes the output locally using a `connection_type` of S3 with a POSIX path argument in `connection_options`, which allows writing to local storage.

```
glueContext.write_dynamic_frame.from_options(\
frame = dyf_splitFields,\
connection_options = {'path': '/home/glue/GlueLocalOutput/'},\
connection_type = 's3',\
format = 'json')
```

DynamicFrameReader class

— methods —

- __init__
- from_rdd
- from_options
- from_catalog

__init__

`__init__(glue_context)`

- `glue_context` – The GlueContext class to use.

from_rdd

`from_rdd(data, name, schema=None, sampleRatio=None)`

Reads a `DynamicFrame` from a Resilient Distributed Dataset (RDD).

- `data` – The dataset to read from.
- `name` – The name to read from.
- `schema` – The schema to read (optional).
- `sampleRatio` – The sample ratio (optional).

from_options

`from_options(connection_type, connection_options={}, format=None, format_options={}, transformation_ctx="")`

Reads a `DynamicFrame` using the specified connection and format.

- `connection_type` – The connection type. Valid values include `s3`, `mysql`, `postgresql`, `redshift`, `sqlserver`, `oracle`, `dynamodb`, and `snowflake`.
- `connection_options` – Connection options, such as path and database table (optional). For more information, see [Connection types and options for ETL in AWS Glue for Spark](). For a `connection_type` of `s3`, Amazon S3 paths are defined in an array.

```
connection_options = {"paths": [ "s3://amzn-s3-demo-bucket/object_a", "s3://amzn-s3-demo-bucket/object_b"]}
```

For JDBC connections, several properties must be defined. Note that the database name must be part of the URL. It can optionally be included in the connection options.

> ⚠ **Warning**
>
> Storing passwords in your script is not recommended. Consider using `boto3` to retrieve them from AWS Secrets Manager or the AWS Glue Data Catalog.

```
connection_options = {"url": "jdbc-url/database", "user": "username",
 "password": passwordVariable,"dbtable": "table-name", "redshiftTmpDir": "s3-tempdir-path"}
```

For a JDBC connection that performs parallel reads, you can set the hashfield option. For example:

```
connection_options = {"url": "jdbc-url/database", "user": "username",
 "password": passwordVariable,"dbtable": "table-name", "redshiftTmpDir": "s3-tempdir-path" , "hashfield": "month"}
```

For more information, see [Reading from JDBC tables in parallel]().

- `format` – A format specification (optional). This is used for an Amazon Simple Storage Service (Amazon S3) or an AWS Glue connection that supports multiple formats. See Data format options for inputs and outputs in AWS Glue for Spark for the formats that are supported.
- `format_options` – Format options for the specified format. See Data format options for inputs and outputs in AWS Glue for Spark for the formats that are supported.
- `transformation_ctx` – The transformation context to use (optional).
- `push_down_predicate` – Filters partitions without having to list and read all the files in your dataset. For more information, see Pre-Filtering Using Pushdown Predicates.

from_catalog

`from_catalog(database, table_name, redshift_tmp_dir="", transformation_ctx="", push_down_predicate="", additional_options={})`

Reads a `DynamicFrame` using the specified catalog namespace and table name.

- `database` – The database to read from.
- `table_name` – The name of the table to read from.
- `redshift_tmp_dir` – An Amazon Redshift temporary directory to use (optional if not reading data from Redshift).
- `transformation_ctx` – The transformation context to use (optional).
- `push_down_predicate` – Filters partitions without having to list and read all the files in your dataset. For more information, see Pre-filtering using pushdown predicates.
- `additional_options` – Additional options provided to AWS Glue.
  - To use a JDBC connection that performs parallel reads, you can set the `hashfield`, `hashexpression`, or `hashpartitions` options. For example:

    ```
    additional_options = {"hashfield": "month"}
    ```

    For more information, see Reading from JDBC tables in parallel.
  - To pass a catalog expression to filter based on the index columns, you can see the `catalogPartitionPredicate` option.

    `catalogPartitionPredicate` — You can pass a catalog expression to filter based on the index columns. This pushes down the filtering to the server side. For more information, see AWS Glue Partition Indexes. Note that `push_down_predicate` and `catalogPartitionPredicate` use different syntaxes. The former one uses Spark SQL standard syntax and the later one uses JSQL parser.

For more information, see Managing partitions for ETL output in AWS Glue.

GlueContext class

Wraps the Apache Spark SparkContext object, and thereby provides mechanisms for interacting with the Apache Spark platform.

__init__

`__init__(sparkContext)`

- `sparkContext` – The Apache Spark context to use.

Creating

- __init__
- getSource
- create_dynamic_frame_from_rdd
- create_dynamic_frame_from_catalog
- create_dynamic_frame_from_options
- create_sample_dynamic_frame_from_catalog
- create_sample_dynamic_frame_from_options
- add_ingestion_time_columns
- create_data_frame_from_catalog
- create_data_frame_from_options
- forEachBatch

getSource

`getSource(connection_type, transformation_ctx = "", **options)`

Creates a `DataSource` object that can be used to read `DynamicFrames` from external sources.

- `connection_type` – The connection type to use, such as Amazon Simple Storage Service (Amazon S3), Amazon Redshift, and JDBC. Valid values include `s3`, `mysql`, `postgresql`, `redshift`, `sqlserver`, `oracle`, and `dynamodb`.
- `transformation_ctx` – The transformation context to use (optional).
- `options` – A collection of optional name-value pairs. For more information, see Connection types and options for ETL in AWS Glue for Spark.

The following is an example of using `getSource`.

```
>>> data_source = context.getSource("file", paths=["/in/path"])
>>> data_source.setFormat("json")
>>> myFrame = data_source.getFrame()
```

create_dynamic_frame_from_rdd

`create_dynamic_frame_from_rdd(data, name, schema=None, sample_ratio=None, transformation_ctx="")`

Returns a `DynamicFrame` that is created from an Apache Spark Resilient Distributed Dataset (RDD).

- `data` – The data source to use.
- `name` – The name of the data to use.
- `schema` – The schema to use (optional).
- `sample_ratio` – The sample ratio to use (optional).
- `transformation_ctx` – The transformation context to use (optional).

create_dynamic_frame_from_catalog

`create_dynamic_frame_from_catalog(database, table_name, redshift_tmp_dir, transformation_ctx = "", push_down_predicate= "", additional_options = {}, catalog_id = None)`

Returns a `DynamicFrame` that is created using a Data Catalog database and table name. When using this method, you provide `format_options` through table properties on the specified AWS Glue Data Catalog table and other options through the `additional_options` argument.

- `Database` – The database to read from.
- `table_name` – The name of the table to read from.
- `redshift_tmp_dir` – An Amazon Redshift temporary directory to use (optional).
- `transformation_ctx` – The transformation context to use (optional).
- `push_down_predicate` – Filters partitions without having to list and read all the files in your dataset. For supported sources and limitations, see Optimizing reads with pushdown in AWS Glue ETL. For more information, see Pre-filtering using pushdown predicates.
- `additional_options` – A collection of optional name-value pairs. The possible options include those listed in Connection types and options for ETL in AWS Glue for Spark except for `endpointUrl`, `streamName`, `bootstrap.servers`, `security.protocol`, `topicName`, `classification`, and `delimiter`. Another supported option is `catalogPartitionPredicate`:

  `catalogPartitionPredicate` — You can pass a catalog expression to filter based on the index columns. This pushes down the filtering to the server side. For more information, see AWS Glue Partition Indexes. Note that `push_down_predicate` and `catalogPartitionPredicate` use different syntaxes. The former one uses Spark SQL standard syntax and the later one uses JSQL parser.
- `catalog_id` — The catalog ID (account ID) of the Data Catalog being accessed. When None, the default account ID of the caller is used.

create_dynamic_frame_from_options

`create_dynamic_frame_from_options(connection_type, connection_options={}, format=None, format_options={}, transformation_ctx = "")`

Returns a `DynamicFrame` created with the specified connection and format.

- `connection_type` – The connection type, such as Amazon S3, Amazon Redshift, and JDBC. Valid values include `s3`, `mysql`, `postgresql`, `redshift`, `sqlserver`, `oracle`, and `dynamodb`.
- `connection_options` – Connection options, such as paths and database table (optional). For a `connection_type` of `s3`, a list of Amazon S3 paths is defined.

```
connection_options = {"paths": ["s3://aws-glue-target/temp"]}
```

For JDBC connections, several properties must be defined. Note that the database name must be part of the URL. It can optionally be included in the connection options.

> ⚠ **Warning**
>
> Storing passwords in your script is not recommended. Consider using `boto3` to retrieve them from AWS Secrets Manager or the AWS Glue Data Catalog.

```
connection_options = {"url": "jdbc-url/database", "user": "username",
 "password": passwordVariable,"dbtable": "table-name", "redshiftTmpDir": "s3-tempdir-
path"}
```

The `dbtable` property is the name of the JDBC table. For JDBC data stores that support schemas within a database, specify `schema.table-name`. If a schema is not provided, then the default "public" schema is used.

For more information, see Connection types and options for ETL in AWS Glue for Spark.

- `format` – A format specification. This is used for an Amazon S3 or an AWS Glue connection that supports multiple formats. See Data format options for inputs and outputs in AWS Glue for Spark for the formats that are supported.
- `format_options` – Format options for the specified format. See Data format options for inputs and outputs in AWS Glue for Spark for the formats that are supported.
- `transformation_ctx` – The transformation context to use (optional).
- `push_down_predicate` – Filters partitions without having to list and read all the files in your dataset. For supported sources and limitations, see Optimizing reads with pushdown in AWS Glue ETL. For more information, see Pre-Filtering Using Pushdown Predicates.

create_sample_dynamic_frame_from_catalog

`create_sample_dynamic_frame_from_catalog(database, table_name, num, redshift_tmp_dir, transformation_ctx = "", push_down_predicate= "", additional_options = {}, sample_options = {}, catalog_id = None)`

Returns a sample `DynamicFrame` that is created using a Data Catalog database and table name. The `DynamicFrame` only contains first num records from a datasource.

- `database` – The database to read from.
- `table_name` – The name of the table to read from.
- `num` – The maximum number of records in the returned sample dynamic frame.
- `redshift_tmp_dir` – An Amazon Redshift temporary directory to use (optional).
- `transformation_ctx` – The transformation context to use (optional).
- `push_down_predicate` – Filters partitions without having to list and read all the files in your dataset. For more information, see Pre-filtering using pushdown predicates.
- `additional_options` – A collection of optional name-value pairs. The possible options include those listed in Connection types and options for ETL in AWS Glue for Spark except for `endpointUrl`, `streamName`, `bootstrap.servers`, `security.protocol`, `topicName`, `classification`, and `delimiter`.
- `sample_options` – Parameters to control sampling behavior (optional). Current available parameters for Amazon S3 sources:
  - `maxSamplePartitions` – The maximum number of partitions the sampling will read. Default value is 10
  - `maxSampleFilesPerPartition` – The maximum number of files the sampling will read in one partition. Default value is 10.

    These parameters help to reduce the time consumed by file listing. For example, suppose the dataset has 1000 partitions, and each partition has 10 files. If you set `maxSamplePartitions` = 10, and `maxSampleFilesPerPartition` = 10, instead of listing all 10,000 files, the sampling will only list and read the first 10 partitions with the first 10 files in each: 10*10 = 100 files in total.
- `catalog_id` – The catalog ID of the Data Catalog being accessed (the account ID of the Data Catalog). Set to `None` by default. `None` defaults to the catalog ID of the calling account in the service.

create_sample_dynamic_frame_from_options

`create_sample_dynamic_frame_from_options(connection_type, connection_options={}, num, sample_options={}, format=None, format_options={}, transformation_ctx = "")`

Returns a sample `DynamicFrame` created with the specified connection and format. The `DynamicFrame` only contains first `num` records from a datasource.

- `connection_type` – The connection type, such as Amazon S3, Amazon Redshift, and JDBC. Valid values include `s3`, `mysql`, `postgresql`, `redshift`, `sqlserver`, `oracle`, and `dynamodb`.
- `connection_options` – Connection options, such as paths and database table (optional). For more information, see Connection types and options for ETL in AWS Glue for Spark.
- `num` – The maximum number of records in the returned sample dynamic frame.
- `sample_options` – Parameters to control sampling behavior (optional). Current available parameters for Amazon S3 sources:
  - `maxSamplePartitions` – The maximum number of partitions the sampling will read. Default value is 10
  - `maxSampleFilesPerPartition` – The maximum number of files the sampling will read in one partition. Default value is 10.

    These parameters help to reduce the time consumed by file listing. For example, suppose the dataset has 1000 partitions, and each partition has 10 files. If you set `maxSamplePartitions` = 10, and `maxSampleFilesPerPartition` = 10, instead of listing all 10,000 files, the sampling will only list and read the first 10 partitions with the first 10 files in each: 10*10 = 100 files in total.
- `format` – A format specification. This is used for an Amazon S3 or an AWS Glue connection that supports multiple formats. See Data format options for inputs and outputs in AWS Glue for Spark for the formats that are supported.
- `format_options` – Format options for the specified format. See Data format options for inputs and outputs in AWS Glue for Spark for the formats that are supported.
- `transformation_ctx` – The transformation context to use (optional).
- `push_down_predicate` – Filters partitions without having to list and read all the files in your dataset. For more information, see Pre-filtering using pushdown predicates.

add_ingestion_time_columns

`add_ingestion_time_columns(dataFrame, timeGranularity = "")`

Appends ingestion time columns like `ingest_year`, `ingest_month`, `ingest_day`, `ingest_hour`, `ingest_minute` to the input `DataFrame`. This function is automatically generated in the script generated by the AWS Glue when you specify a Data Catalog table with Amazon S3 as the target. This function automatically updates the partition with ingestion time columns on the output table. This allows the output data to be automatically partitioned on ingestion time without requiring explicit ingestion time columns in the input data.

- `dataFrame` – The `dataFrame` to append the ingestion time columns to.
- `timeGranularity` – The granularity of the time columns. Valid values are "`day`", "`hour`" and "`minute`". For example, if "`hour`" is passed in to the function, the original `dataFrame` will have "`ingest_year`", "`ingest_month`", "`ingest_day`", and "`ingest_hour`" time columns appended.

Returns the data frame after appending the time granularity columns.

Example:

```
dynamic_frame = DynamicFrame.fromDF(glueContext.add_ingestion_time_columns(dataFrame,
 "hour"))
```

create_data_frame_from_catalog

`create_data_frame_from_catalog(database, table_name, transformation_ctx = "", additional_options = {})`

Returns a `DataFrame` that is created using information from a Data Catalog table.

- `database` – The Data Catalog database to read from.
- `table_name` – The name of the Data Catalog table to read from.
- `transformation_ctx` – The transformation context to use (optional).
- `additional_options` – A collection of optional name-value pairs. The possible options include those listed in Connection types and options for ETL in AWS Glue for Spark for streaming sources, such as `startingPosition`, `maxFetchTimeInMs`, and `startingOffsets`.
  - `useSparkDataSource` – When set to true, forces AWS Glue to use the native Spark Data Source API to read the table. The Spark Data Source API supports the following formats: AVRO, binary, CSV, JSON, ORC, Parquet, and text. In a Data Catalog table, you specify the format using the `classification` property. To learn more about the Spark Data Source API, see the official Apache Spark documentation.

    Using `create_data_frame_from_catalog` with `useSparkDataSource` has the following benefits:

  - Directly returns a `DataFrame` and provides an alternative to `create_dynamic_frame.from_catalog().toDF()`.
  - Supports AWS Lake Formation table-level permission control for native formats.
  - Supports reading data lake formats without AWS Lake Formation table-level permission control. For more information, see Using data lake frameworks with AWS Glue ETL jobs.

  When you enable `useSparkDataSource`, you can also add any of the Spark Data Source options in `additional_options` as needed. AWS Glue passes these options directly to the Spark reader.

- `useCatalogSchema` – When set to true, AWS Glue applies the Data Catalog schema to the resulting `DataFrame`. Otherwise, the reader infers the schema from the data. When you enable `useCatalogSchema`, you must also set `useSparkDataSource` to true.

Limitations

Consider the following limitations when you use the `useSparkDataSource` option:

- When you use `useSparkDataSource`, AWS Glue creates a new `DataFrame` in a separate Spark session that is different from the original Spark session.
- Spark DataFrame partition filtering doesn't work with the following AWS Glue features.
  - Job bookmarks
  - Excluding Amazon S3 storage classes
  - Catalog partition predicates

  To use partition filtering with these features, you can use the AWS Glue pushdown predicate. For more information, see Pre-filtering using pushdown predicates. Filtering on non-partitioned columns is not affected.

  The following example script demonstrates the incorrect way to perform partition filtering with the `excludeStorageClasses` option.

```
// Incorrect partition filtering using Spark filter with excludeStorageClasses
read_df = glueContext.create_data_frame.from_catalog(
    database=database_name,
    table_name=table_name,
    additional_options = {
      "useSparkDataSource": True,
        "excludeStorageClasses" : ["GLACIER", "DEEP_ARCHIVE"]
    }
)

//  Suppose year and month are partition keys.
//  Filtering on year and month won't work, the filtered_df will still
//  contain data with other year/month values.
filtered_df = read_df.filter("year == '2017 and month == '04' and 'state == 'CA'")
```

The following example script demonstrates the correct way to use a pushdown predicate in order to perform partition filtering with the `excludeStorageClasses` option.

```
// Correct partition filtering using the AWS Glue pushdown predicate
// with excludeStorageClasses
read_df = glueContext.create_data_frame.from_catalog(
    database=database_name,
    table_name=table_name,
    //  Use AWS Glue pushdown predicate to perform partition filtering
    push_down_predicate = "(year=='2017' and month=='04')"
    additional_options = {
      "useSparkDataSource": True,
      "excludeStorageClasses" : ["GLACIER", "DEEP_ARCHIVE"]
    }
)

//  Use Spark filter only on non-partitioned columns
filtered_df = read_df.filter("state == 'CA'")
```

Example: Creating a CSV table using the Spark data source reader

```
//  Read a CSV table with '\t' as separator
read_df = glueContext.create_data_frame.from_catalog(
    database=<database_name>,
    table_name=<table_name>,
    additional_options = {"useSparkDataSource": True,  "sep": '\t'}
)
```

create_data_frame_from_options

`create_data_frame_from_options(connection_type, connection_options={}, format=None, format_options={}, transformation_ctx = "")`

This API is now deprecated. Instead use the `getSource()` API. Returns a `DataFrame` created with the specified connection and format. Use this function only with AWS Glue streaming sources.

- `connection_type` – The streaming connection type. Valid values include `kinesis` and `kafka`.
- `connection_options` – Connection options, which are different for Kinesis and Kafka. You can find the list of all connection options for each streaming data source at Connection types and options for ETL in AWS Glue for Spark. Note the following differences in streaming connection options:
  - Kinesis streaming sources require `streamARN`, `startingPosition`, `inferSchema`, and `classification`.
  - Kafka streaming sources require `connectionName`, `topicName`, `startingOffsets`, `inferSchema`, and `classification`.
- `format` – A format specification. This is used for an Amazon S3 or an AWS Glue connection that supports multiple formats. For information about the supported formats, see Data format options for inputs and outputs in AWS Glue for Spark.
- `format_options` – Format options for the specified format. For information about the supported format options, see Data format options for inputs and outputs in AWS Glue for Spark.
- `transformation_ctx` – The transformation context to use (optional).

Example for Amazon Kinesis streaming source:

```
kinesis_options =
   { "streamARN": "arn:aws:kinesis:us-east-2:777788889999:stream/fromOptionsStream",
     "startingPosition": "TRIM_HORIZON",
     "inferSchema": "true",
     "classification": "json"
   }
data_frame_datasource0 =
 glueContext.create_data_frame.from_options(connection_type="kinesis",
 connection_options=kinesis_options)
```

Example for Kafka streaming source:

```
kafka_options =
    { "connectionName": "ConfluentKafka",
      "topicName": "kafka-auth-topic",
      "startingOffsets": "earliest",
      "inferSchema": "true",
      "classification": "json"
    }
data_frame_datasource0 =
 glueContext.create_data_frame.from_options(connection_type="kafka",
 connection_options=kafka_options)
```

forEachBatch

`forEachBatch(frame, batch_function, options)`

Applies the `batch_function` passed in to every micro batch that is read from the Streaming source.

- `frame` – The DataFrame containing the current micro batch.
- `batch_function` – A function that will be applied for every micro batch.
- `options` – A collection of key-value pairs that holds information about how to process micro batches. The following options are required:
  - `windowSize` – The amount of time to spend processing each batch.
  - `checkpointLocation` – The location where checkpoints are stored for the streaming ETL job.
  - `batchMaxRetries` – The maximum number of times to retry the batch if it fails. The default value is 3. This option is only configurable for Glue version 2.0 and above.

Example:

```
glueContext.forEachBatch(
    frame = data_frame_datasource0,
    batch_function = processBatch,
    options = {
        "windowSize": "100 seconds",
        "checkpointLocation": "s3://kafka-auth-dataplane/confluent-test/output/
checkpoint/"
    }
)
```

```
def processBatch(data_frame, batchId):
    if (data_frame.count() > 0):
        datasource0 = DynamicFrame.fromDF(
          glueContext.add_ingestion_time_columns(data_frame, "hour"),
          glueContext, "from_data_frame"
        )
        additionalOptions_datasink1 = {"enableUpdateCatalog": True}
        additionalOptions_datasink1["partitionKeys"] = ["ingest_yr", "ingest_mo",
 "ingest_day"]
        datasink1 = glueContext.write_dynamic_frame.from_catalog(
          frame = datasource0,
          database = "tempdb",
          table_name = "kafka-auth-table-output",
          transformation_ctx = "datasink1",
          additional_options = additionalOptions_datasink1
        )
```

Working with datasets in Amazon S3

- purge_table
- purge_s3_path
- transition_table
- transition_s3_path

purge_table

`purge_table(catalog_id=None, database="", table_name="", options={}, transformation_ctx="")`

Deletes files from Amazon S3 for the specified catalog's database and table. If all files in a partition are deleted, that partition is also deleted from the catalog. We don't support purge_table action on tables registered with Lake Formation.

If you want to be able to recover deleted objects, you can turn on object versioning on the Amazon S3 bucket. When an object is deleted from a bucket that doesn't have object versioning enabled, the object can't be recovered. For more information about how to recover deleted objects in a version-enabled bucket, see How can I retrieve an Amazon S3 object that was deleted? in the AWS Support Knowledge Center.

- `catalog_id` – The catalog ID of the Data Catalog being accessed (the account ID of the Data Catalog). Set to None by default. None defaults to the catalog ID of the calling account in the service.
- `database` – The database to use.
- `table_name` – The name of the table to use.
- `options` – Options to filter files to be deleted and for manifest file generation.
  - `retentionPeriod` – Specifies a period in number of hours to retain files. Files newer than the retention period are retained. Set to 168 hours (7 days) by default.
  - `partitionPredicate` – Partitions satisfying this predicate are deleted. Files within the retention period in these partitions are not deleted. Set to `""` – empty by default.
  - `excludeStorageClasses` – Files with storage class in the `excludeStorageClasses` set are not deleted. The default is `Set()` – an empty set.
  - `manifestFilePath` – An optional path for manifest file generation. All files that were successfully purged are recorded in `Success.csv`, and those that failed in `Failed.csv`
- `transformation_ctx` – The transformation context to use (optional). Used in the manifest file path.

Example

```
glueContext.purge_table("database", "table", {"partitionPredicate": "(month=='march')",
 "retentionPeriod": 1, "excludeStorageClasses": ["STANDARD_IA"], "manifestFilePath":
 "s3://bucketmanifest/"})
```

purge_s3_path

`purge_s3_path(s3_path, options={}, transformation_ctx="")`

Deletes files from the specified Amazon S3 path recursively.

If you want to be able to recover deleted objects, you can turn on object versioning on the Amazon S3 bucket. When an object is deleted from a bucket that doesn't have object versioning turned on, the object can't be recovered. For more information about how to recover deleted objects in a bucket with versioning, see How can I retrieve an Amazon S3 object that was deleted? in the Support Knowledge Center.

- `s3_path` – The path in Amazon S3 of the files to be deleted in the format `s3://<bucket>/<prefix>/`

- `options` – Options to filter files to be deleted and for manifest file generation.
  - `retentionPeriod` – Specifies a period in number of hours to retain files. Files newer than the retention period are retained. Set to 168 hours (7 days) by default.
  - `excludeStorageClasses` – Files with storage class in the `excludeStorageClasses` set are not deleted. The default is `Set()` – an empty set.
  - `manifestFilePath` – An optional path for manifest file generation. All files that were successfully purged are recorded in `Success.csv`, and those that failed in `Failed.csv`
- `transformation_ctx` – The transformation context to use (optional). Used in the manifest file path.

Example

```
glueContext.purge_s3_path("s3://bucket/path/", {"retentionPeriod": 1,
 "excludeStorageClasses": ["STANDARD_IA"], "manifestFilePath": "s3://bucketmanifest/"})
```

transition_table

`transition_table(database, table_name, transition_to, options={}, transformation_ctx="", catalog_id=None)`

Transitions the storage class of the files stored on Amazon S3 for the specified catalog's database and table.

You can transition between any two storage classes. For the `GLACIER` and `DEEP_ARCHIVE` storage classes, you can transition to these classes. However, you would use an `S3 RESTORE` to transition from `GLACIER` and `DEEP_ARCHIVE` storage classes.

If you're running AWS Glue ETL jobs that read files or partitions from Amazon S3, you can exclude some Amazon S3 storage class types. For more information, see Excluding Amazon S3 Storage Classes.

- `database` – The database to use.
- `table_name` – The name of the table to use.
- `transition_to` – The Amazon S3 storage class to transition to.
- `options` – Options to filter files to be deleted and for manifest file generation.
  - `retentionPeriod` – Specifies a period in number of hours to retain files. Files newer than the retention period are retained. Set to 168 hours (7 days) by default.

  - `partitionPredicate` – Partitions satisfying this predicate are transitioned. Files within the retention period in these partitions are not transitioned. Set to `""` – empty by default.
  - `excludeStorageClasses` – Files with storage class in the `excludeStorageClasses` set are not transitioned. The default is `Set()` – an empty set.
  - `manifestFilePath` – An optional path for manifest file generation. All files that were successfully transitioned are recorded in `Success.csv`, and those that failed in `Failed.csv`
  - `accountId` – The Amazon Web Services account ID to run the transition transform. Mandatory for this transform.
  - `roleArn` – The AWS role to run the transition transform. Mandatory for this transform.
- `transformation_ctx` – The transformation context to use (optional). Used in the manifest file path.
- `catalog_id` – The catalog ID of the Data Catalog being accessed (the account ID of the Data Catalog). Set to `None` by default. `None` defaults to the catalog ID of the calling account in the service.

Example

```
glueContext.transition_table("database", "table", "STANDARD_IA", {"retentionPeriod": 1,
 "excludeStorageClasses": ["STANDARD_IA"], "manifestFilePath": "s3://bucketmanifest/",
 "accountId": "12345678901", "roleArn": "arn:aws:iam::123456789012:user/example-
username"})
```

transition_s3_path

`transition_s3_path(s3_path, transition_to, options={}, transformation_ctx="")`

Transitions the storage class of the files in the specified Amazon S3 path recursively.

You can transition between any two storage classes. For the `GLACIER` and `DEEP_ARCHIVE` storage classes, you can transition to these classes. However, you would use an `S3 RESTORE` to transition from `GLACIER` and `DEEP_ARCHIVE` storage classes.

If you're running AWS Glue ETL jobs that read files or partitions from Amazon S3, you can exclude some Amazon S3 storage class types. For more information, see Excluding Amazon S3 Storage Classes.

- `s3_path` – The path in Amazon S3 of the files to be transitioned in the format `s3://<bucket>/<prefix>/`
- `transition_to` – The Amazon S3 storage class to transition to.
- `options` – Options to filter files to be deleted and for manifest file generation.
  - `retentionPeriod` – Specifies a period in number of hours to retain files. Files newer than the retention period are retained. Set to 168 hours (7 days) by default.
  - `partitionPredicate` – Partitions satisfying this predicate are transitioned. Files within the retention period in these partitions are not transitioned. Set to `""` – empty by default.
  - `excludeStorageClasses` – Files with storage class in the `excludeStorageClasses` set are not transitioned. The default is `Set()` – an empty set.
  - `manifestFilePath` – An optional path for manifest file generation. All files that were successfully transitioned are recorded in `Success.csv`, and those that failed in `Failed.csv`
  - `accountId` – The Amazon Web Services account ID to run the transition transform. Mandatory for this transform.
  - `roleArn` – The AWS role to run the transition transform. Mandatory for this transform.
- `transformation_ctx` – The transformation context to use (optional). Used in the manifest file path.

Example

```
glueContext.transition_s3_path("s3://bucket/prefix/", "STANDARD_IA",
 {"retentionPeriod": 1, "excludeStorageClasses": ["STANDARD_IA"],
 "manifestFilePath": "s3://bucketmanifest/", "accountId": "12345678901", "roleArn":
 "arn:aws:iam::123456789012:user/example-username"})
```

Extracting

- extract_jdbc_conf

extract_jdbc_conf

`extract_jdbc_conf(connection_name, catalog_id = None)`

Returns a `dict` with keys with the configuration properties from the AWS Glue connection object in the Data Catalog.

- `user` – The database user name.
- `password` – The database password.
- `vendor` – Specifies a vendor (`mysql`, `postgresql`, `oracle`, `sqlserver`, etc.).
- `enforceSSL` – A boolean string indicating if a secure connection is required.
- `customJDBCCert` – Use a specific client certificate from the Amazon S3 path indicated.
- `skipCustomJDBCCertValidation` – A boolean string indicating if the `customJDBCCert` must be validated by a CA.
- `customJDBCCertString` – Additional information about the custom certificate, specific for the driver type.
- `url` – (Deprecated) JDBC URL with only protocol, server and port.
- `fullUrl` – JDBC URL as entered when the connection was created (Available in AWS Glue version 3.0 or later).

Example retrieving JDBC configurations:

```
jdbc_conf = glueContext.extract_jdbc_conf(connection_name="your_glue_connection_name")
print(jdbc_conf)
>>> {'enforceSSL': 'false', 'skipCustomJDBCCertValidation': 'false', 'url':
 'jdbc:mysql://myserver:3306', 'fullUrl': 'jdbc:mysql://myserver:3306/mydb',
 'customJDBCCertString': '', 'user': 'admin', 'customJDBCCert': '', 'password': '1234',
 'vendor': 'mysql'}
```

Transactions

- start_transaction
- commit_transaction
- cancel_transaction

start_transaction

`start_transaction(read_only)`

Start a new transaction. Internally calls the Lake Formation startTransaction API.

- `read_only` – (Boolean) Indicates whether this transaction should be read only or read and write. Writes made using a read-only transaction ID will be rejected. Read-only transactions do not need to be committed.

Returns the transaction ID.

commit_transaction

`commit_transaction(transaction_id, wait_for_commit = True)`

Attempts to commit the specified transaction. `commit_transaction` may return before the transaction has finished committing. Internally calls the Lake Formation commitTransaction API.

- `transaction_id` – (String) The transaction to commit.
- `wait_for_commit` – (Boolean) Determines whether the `commit_transaction` returns immediately. The default value is true. If false, `commit_transaction` polls and waits until the transaction is committed. The amount of wait time is restricted to 1 minute using exponential backoff with a maximum of 6 retry attempts.

Returns a Boolean to indicate whether the commit is done or not.

cancel_transaction

`cancel_transaction(transaction_id)`

Attempts to cancel the specified transaction. Returns a `TransactionCommittedException` exception if the transaction was previously committed. Internally calls the Lake Formation CancelTransaction API.

- `transaction_id` – (String) The transaction to cancel.

Writing

- getSink
- write_dynamic_frame_from_options
- write_from_options
- write_dynamic_frame_from_catalog

- write_data_frame_from_catalog
- write_dynamic_frame_from_jdbc_conf
- write_from_jdbc_conf

getSink

`getSink(connection_type, format = None, transformation_ctx = "", **options)`

Gets a `DataSink` object that can be used to write `DynamicFrames` to external sources. Check the SparkSQL `format` first to be sure to get the expected sink.

- `connection_type` – The connection type to use, such as Amazon S3, Amazon Redshift, and JDBC. Valid values include `s3`, `mysql`, `postgresql`, `redshift`, `sqlserver`, `oracle`, `kinesis`, and `kafka`.
- `format` – The SparkSQL format to use (optional).
- `transformation_ctx` – The transformation context to use (optional).
- `options` – A collection of name-value pairs used to specify the connection options. Some of the possible values are:
  - `user` and `password`: For authorization
  - `url`: The endpoint for the data store
  - `dbtable`: The name of the target table
  - `bulkSize`: Degree of parallelism for insert operations

The options that you can specify depends on the connection type. See Connection types and options for ETL in AWS Glue for Spark for additional values and examples.

Example:

```
>>> data_sink = context.getSink("s3")
>>> data_sink.setFormat("json"),
>>> data_sink.writeFrame(myFrame)
```

write_dynamic_frame_from_options

`write_dynamic_frame_from_options(frame, connection_type, connection_options={}, format=None, format_options={}, transformation_ctx = "")`

Writes and returns a `DynamicFrame` using the specified connection and format.

- `frame` – The `DynamicFrame` to write.
- `connection_type` – The connection type, such as Amazon S3, Amazon Redshift, and JDBC. Valid values include `s3`, `mysql`, `postgresql`, `redshift`, `sqlserver`, `oracle`, `kinesis`, and `kafka`.
- `connection_options` – Connection options, such as path and database table (optional). For a `connection_type` of `s3`, an Amazon S3 path is defined.

  ```
  connection_options = {"path": "s3://aws-glue-target/temp"}
  ```

  For JDBC connections, several properties must be defined. Note that the database name must be part of the URL. It can optionally be included in the connection options.

  > **Warning**
  >
  > Storing passwords in your script is not recommended. Consider using `boto3` to retrieve them from AWS Secrets Manager or the AWS Glue Data Catalog.

  ```
  connection_options = {"url": "jdbc-url/database", "user": "username",
   "password": passwordVariable,"dbtable": "table-name", "redshiftTmpDir": "s3-tempdir-path"}
  ```

  The `dbtable` property is the name of the JDBC table. For JDBC data stores that support schemas within a database, specify `schema.table-name`. If a schema is not provided, then the default "public" schema is used.

  For more information, see Connection types and options for ETL in AWS Glue for Spark.
- `format` – A format specification. This is used for an Amazon S3 or an AWS Glue connection that supports multiple formats. See Data format options for inputs and outputs in AWS Glue for Spark for the formats that are supported.

- `format_options` – Format options for the specified format. See Data format options for inputs and outputs in AWS Glue for Spark for the formats that are supported.
- `transformation_ctx` – A transformation context to use (optional).

write_from_options

`write_from_options(frame_or_dfc, connection_type, connection_options={}, format={}, format_options={}, transformation_ctx = "")`

Writes and returns a `DynamicFrame` or `DynamicFrameCollection` that is created with the specified connection and format information.

- `frame_or_dfc` – The `DynamicFrame` or `DynamicFrameCollection` to write.
- `connection_type` – The connection type, such as Amazon S3, Amazon Redshift, and JDBC. Valid values include `s3`, `mysql`, `postgresql`, `redshift`, `sqlserver`, and `oracle`.
- `connection_options` – Connection options, such as path and database table (optional). For a `connection_type` of `s3`, an Amazon S3 path is defined.

  ```
  connection_options = {"path": "s3://aws-glue-target/temp"}
  ```

  For JDBC connections, several properties must be defined. Note that the database name must be part of the URL. It can optionally be included in the connection options.

  > **Warning**
  >
  > Storing passwords in your script is not recommended. Consider using `boto3` to retrieve them from AWS Secrets Manager or the AWS Glue Data Catalog.

  ```
  connection_options = {"url": "jdbc-url/database", "user": "username",
   "password": passwordVariable,"dbtable": "table-name", "redshiftTmpDir": "s3-tempdir-path"}
  ```

  The `dbtable` property is the name of the JDBC table. For JDBC data stores that support schemas within a database, specify `schema.table-name`. If a schema is not provided, then the default "public" schema is used.

For more information, see Connection types and options for ETL in AWS Glue for Spark.

- `format` – A format specification. This is used for an Amazon S3 or an AWS Glue connection that supports multiple formats. See Data format options for inputs and outputs in AWS Glue for Spark for the formats that are supported.
- `format_options` – Format options for the specified format. See Data format options for inputs and outputs in AWS Glue for Spark for the formats that are supported.
- `transformation_ctx` – A transformation context to use (optional).

write_dynamic_frame_from_catalog

`write_dynamic_frame_from_catalog(frame, database, table_name, redshift_tmp_dir, transformation_ctx = "", additional_options = {}, catalog_id = None)`

Writes and returns a `DynamicFrame` using information from a Data Catalog database and table.

- `frame` – The `DynamicFrame` to write.
- `Database` – The Data Catalog database that contains the table.
- `table_name` – The name of the Data Catalog table associated with the target.
- `redshift_tmp_dir` – An Amazon Redshift temporary directory to use (optional).
- `transformation_ctx` – The transformation context to use (optional).
- `additional_options` – A collection of optional name-value pairs.
- `catalog_id` — The catalog ID (account ID) of the Data Catalog being accessed. When None, the default account ID of the caller is used.

write_data_frame_from_catalog

`write_data_frame_from_catalog(frame, database, table_name, redshift_tmp_dir, transformation_ctx = "", additional_options = {}, catalog_id = None)`

Writes and returns a `DataFrame` using information from a Data Catalog database and table. This method supports writing to data lake formats (Hudi, Iceberg, and Delta Lake). For more information, see Using data lake frameworks with AWS Glue ETL jobs.

- `frame` – The `DataFrame` to write.
- `Database` – The Data Catalog database that contains the table.
- `table_name` – The name of the Data Catalog table that is associated with the target.
- `redshift_tmp_dir` – An Amazon Redshift temporary directory to use (optional).
- `transformation_ctx` – The transformation context to use (optional).
- `additional_options` – A collection of optional name-value pairs.
  - `useSparkDataSink` – When set to true, forces AWS Glue to use the native Spark Data Sink API to write to the table. When you enable this option, you can add any Spark Data Source options to `additional_options` as needed. AWS Glue passes these options directly to the Spark writer.
- `catalog_id` – The catalog ID (account ID) of the Data Catalog being accessed. When you don't specify a value, the default account ID of the caller is used.

Limitations

Consider the following limitations when you use the `useSparkDataSink` option:

- The `enableUpdateCatalog` option isn't supported when you use the `useSparkDataSink` option.

Example: Writing to a Hudi table using the Spark Data Source writer

```
hudi_options = {
    'useSparkDataSink': True,
    'hoodie.table.name': <table_name>,
    'hoodie.datasource.write.storage.type': 'COPY_ON_WRITE',
    'hoodie.datasource.write.recordkey.field': 'product_id',
    'hoodie.datasource.write.table.name': <table_name>,
    'hoodie.datasource.write.operation': 'upsert',
    'hoodie.datasource.write.precombine.field': 'updated_at',
    'hoodie.datasource.write.hive_style_partitioning': 'true',
    'hoodie.upsert.shuffle.parallelism': 2,
    'hoodie.insert.shuffle.parallelism': 2,
    'hoodie.datasource.hive_sync.enable': 'true',
    'hoodie.datasource.hive_sync.database': <database_name>,
    'hoodie.datasource.hive_sync.table': <table_name>,
    'hoodie.datasource.hive_sync.use_jdbc': 'false',
    'hoodie.datasource.hive_sync.mode': 'hms'}
```

```
glueContext.write_data_frame.from_catalog(
    frame = <df_product_inserts>,
    database = <database_name>,
    table_name = <table_name>,
    additional_options = hudi_options
)
```

write_dynamic_frame_from_jdbc_conf

`write_dynamic_frame_from_jdbc_conf(frame, catalog_connection, connection_options={}, redshift_tmp_dir = "", transformation_ctx = "", catalog_id = None)`

Writes and returns a `DynamicFrame` using the specified JDBC connection information.

- `frame` – The `DynamicFrame` to write.
- `catalog_connection` – A catalog connection to use.
- `connection_options` – Connection options, such as path and database table (optional). For more information, see Connection types and options for ETL in AWS Glue for Spark.
- `redshift_tmp_dir` – An Amazon Redshift temporary directory to use (optional).
- `transformation_ctx` – A transformation context to use (optional).
- `catalog_id` — The catalog ID (account ID) of the Data Catalog being accessed. When None, the default account ID of the caller is used.

write_from_jdbc_conf

`write_from_jdbc_conf(frame_or_dfc, catalog_connection, connection_options={}, redshift_tmp_dir = "", transformation_ctx = "", catalog_id = None)`

Writes and returns a `DynamicFrame` or `DynamicFrameCollection` using the specified JDBC connection information.

- `frame_or_dfc` – The `DynamicFrame` or `DynamicFrameCollection` to write.
- `catalog_connection` – A catalog connection to use.
- `connection_options` – Connection options, such as path and database table (optional). For more information, see Connection types and options for ETL in AWS Glue for Spark.

- `redshift_tmp_dir` – An Amazon Redshift temporary directory to use (optional).
- `transformation_ctx` – A transformation context to use (optional).
- `catalog_id` — The catalog ID (account ID) of the Data Catalog being accessed. When None, the default account ID of the caller is used.

AWS Glue PySpark transforms reference

AWS Glue provides the following built-in transforms that you can use in PySpark ETL operations. Your data passes from transform to transform in a data structure called a *DynamicFrame*, which is an extension to an Apache Spark SQL `DataFrame`. The `DynamicFrame` contains your data, and you reference its schema to process your data.

Most of these transforms also exist as methods of the `DynamicFrame` class. For more information, see DynamicFrame transforms .

- GlueTransform base class
- ApplyMapping class
- DropFields class
- DropNullFields class
- ErrorsAsDynamicFrame class
- EvaluateDataQuality class
- FillMissingValues class
- Filter class
- FindIncrementalMatches class
- FindMatches class
- FlatMap class
- Join class
- Map class
- MapToCollection class
- mergeDynamicFrame
- Relationalize class
- RenameField class
- ResolveChoice class

- SelectFields class
- SelectFromCollection class
- Simplify_ddb_json class
- Spigot class
- SplitFields class
- SplitRows class
- Unbox class
- UnnestFrame class

GlueTransform base class

The base class that all the `awsglue.transforms` classes inherit from.

The classes all define a `__call__` method. They either override the `GlueTransform` class methods listed in the following sections, or they are called using the class name by default.

Methods

- apply(cls, *args, **kwargs)
- name(cls)
- describeArgs(cls)
- describeReturn(cls)
- describeTransform(cls)
- describeErrors(cls)
- describe(cls)

apply(cls, *args, **kwargs)

Applies the transform by calling the transform class, and returns the result.

- `cls` – The `self` class object.

name(cls)

Returns the name of the derived transform class.

- `cls` – The `self` class object.

describeArgs(cls)

- `cls` – The `self` class object.

Returns a list of dictionaries, each corresponding to a named argument, in the following format:

```
[
  {
    "name": "(name of argument)",
    "type": "(type of argument)",
    "description": "(description of argument)",
    "optional": "(Boolean, True if the argument is optional)",
    "defaultValue": "(Default value string, or None)(String; the default value, or
 None)"
  },
...
]
```

Raises a `NotImplementedError` exception when called in a derived transform where it is not implemented.

describeReturn(cls)

- `cls` – The `self` class object.

Returns a dictionary with information about the return type, in the following format:

```
{
  "type": "(return type)",
  "description": "(description of output)"
}
```

Raises a `NotImplementedError` exception when called in a derived transform where it is not implemented.

describeTransform(cls)

Returns a string describing the transform.

- `cls` – The `self` class object.

Raises a `NotImplementedError` exception when called in a derived transform where it is not implemented.

describeErrors(cls)

- `cls` – The `self` class object.

Returns a list of dictionaries, each describing a possible exception thrown by this transform, in the following format:

```
[
  {
    "type": "(type of error)",
    "description": "(description of error)"
  },
...
]
```

describe(cls)

- `cls` – The `self` class object.

Returns an object with the following format:

```
{
  "transform" : {
    "name" : cls.name( ),
    "args" : cls.describeArgs( ),
    "returns" : cls.describeReturn( ),
    "raises" : cls.describeErrors( ),
    "location" : "internal"
  }
}
```

ApplyMapping class

Applies a mapping in a `DynamicFrame`.

Example

We recommend that you use the `DynamicFrame.apply_mapping()` method to apply a mapping in a `DynamicFrame`. To view a code example, see Example: Use apply_mapping to rename fields and change field types.

Methods

- __call__
- apply
- name
- describeArgs
- describeReturn
- describeTransform
- describeErrors
- Describe

__call__(frame, mappings, transformation_ctx = "", info = "", stageThreshold = 0, totalThreshold = 0)

Applies a declarative mapping to a specified `DynamicFrame`.

- `frame` – The `DynamicFrame` to apply the mapping to (required).
- `mappings` – A list of mapping tuples (required). Each consists of: (source column, source type, target column, target type).

  If the source column has a dot "`.`" in the name, you must place back-ticks "`` ` ` ``" around it. For example, to map `this.old.name` (string) to `thisNewName`, you would use the following tuple:

  ```
  ("`this.old.name`", "string", "thisNewName", "string")
  ```
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string that is associated with errors in the transformation (optional).
- `stageThreshold` – The maximum number of errors that can occur in the transformation before it errors out (optional). The default is zero.
- `totalThreshold` – The maximum number of errors that can occur overall before processing errors out (optional). The default is zero.

Returns only the fields of the `DynamicFrame` that are specified in the "mapping" tuples.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

DropFields class

Drops fields within a `DynamicFrame`.

Example

We recommend that you use the `DynamicFrame.drop_fields()` method to drop fields from a `DynamicFrame`. To view a code example, see Example: Use drop_fields to remove fields from a `DynamicFrame`.

Methods

- __call__
- apply

- name
- describeArgs
- describeReturn
- describeTransform
- describeErrors
- Describe

__call__(frame, paths, transformation_ctx = "", info = "", stageThreshold = 0, totalThreshold = 0)

Drops nodes within a `DynamicFrame`.

- `frame` – The `DynamicFrame` to drop the nodes in (required).
- `paths` – A list of full paths to the nodes to drop (required).
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string associated with errors in the transformation (optional).
- `stageThreshold` – The maximum number of errors that can occur in the transformation before it errors out (optional). The default is zero.
- `totalThreshold` – The maximum number of errors that can occur overall before processing errors out (optional). The default is zero.

Returns a new `DynamicFrame` without the specified fields.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

DropNullFields class

Drops all null fields in a `DynamicFrame` whose type is `NullType`. These are fields with missing or null values in every record in the `DynamicFrame` dataset.

Example

This example uses `DropNullFields` to create a new `DynamicFrame` where fields of type `NullType` have been dropped. In order to demonstrate `DropNullFields`, we add a new column named `empty_column` with type null to the already-loaded `persons` dataset.

> **Note**
>
> To access the dataset that is used in this example, see Code example: Joining and relationalizing data and follow the instructions in Step 1: Crawl the data in the Amazon S3 bucket.

```
# Example: Use DropNullFields to create a new DynamicFrame without NullType fields

from pyspark.context import SparkContext
from awsglue.context import GlueContext
from pyspark.sql.functions import lit
from pyspark.sql.types import NullType
from awsglue.dynamicframe import DynamicFrame
from awsglue.transforms import DropNullFields

# Create GlueContext
sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)
```

```
# Create DynamicFrame
persons = glueContext.create_dynamic_frame.from_catalog(
    database="legislators", table_name="persons_json"
)
print("Schema for the persons DynamicFrame:")
persons.printSchema()

# Add new column "empty_column" with NullType
persons_with_nulls = persons.toDF().withColumn("empty_column",
 lit(None).cast(NullType()))
persons_with_nulls_dyf = DynamicFrame.fromDF(persons_with_nulls, glueContext,
 "persons_with_nulls")
print("Schema for the persons_with_nulls_dyf DynamicFrame:")
persons_with_nulls_dyf.printSchema()

# Remove the NullType field
persons_no_nulls = DropNullFields.apply(persons_with_nulls_dyf)
print("Schema for the persons_no_nulls DynamicFrame:")
persons_no_nulls.printSchema()
```

Output

```
Schema for the persons DynamicFrame:
root
|-- family_name: string
|-- name: string
|-- links: array
|    |-- element: struct
|    |    |-- note: string
|    |    |-- url: string
|-- gender: string
|-- image: string
|-- identifiers: array
|    |-- element: struct
|    |    |-- scheme: string
|    |    |-- identifier: string
|-- other_names: array
|    |-- element: struct
|    |    |-- lang: string
|    |    |-- note: string
|    |    |-- name: string
|-- sort_name: string
|-- images: array
|    |-- element: struct
|    |    |-- url: string
|-- given_name: string
|-- birth_date: string
|-- id: string
|-- contact_details: array
|    |-- element: struct
|    |    |-- type: string
|    |    |-- value: string
|-- death_date: string

Schema for the persons_with_nulls_dyf DynamicFrame:
root
|-- family_name: string
|-- name: string
|-- links: array
|    |-- element: struct
|    |    |-- note: string
|    |    |-- url: string
|-- gender: string
|-- image: string
|-- identifiers: array
|    |-- element: struct
|    |    |-- scheme: string
|    |    |-- identifier: string
|-- other_names: array
|    |-- element: struct
|    |    |-- lang: string
|    |    |-- note: string
|    |    |-- name: string
|-- sort_name: string
|-- images: array
|    |-- element: struct
|    |    |-- url: string
|-- given_name: string
|-- birth_date: string
|-- id: string
|-- contact_details: array
|    |-- element: struct
|    |    |-- type: string
|    |    |-- value: string
|-- death_date: string
|-- empty_column: null
```

```
null_fields ['empty_column']
Schema for the persons_no_nulls DynamicFrame:
root
|-- family_name: string
|-- name: string
|-- links: array
|    |-- element: struct
|    |    |-- note: string
|    |    |-- url: string
|-- gender: string
|-- image: string
|-- identifiers: array
|    |-- element: struct
|    |    |-- scheme: string
|    |    |-- identifier: string
|-- other_names: array
|    |-- element: struct
|    |    |-- lang: string
|    |    |-- note: string
|    |    |-- name: string
|-- sort_name: string
|-- images: array
|    |-- element: struct
|    |    |-- url: string
|-- given_name: string
|-- birth_date: string
|-- id: string
|-- contact_details: array
|    |-- element: struct
|    |    |-- type: string
|    |    |-- value: string
|-- death_date: string
```

Methods

- __call__
- apply
- name
- describeArgs
- describeReturn

- describeTransform
- describeErrors
- Describe

__call__(frame, transformation_ctx = "", info = "", stageThreshold = 0, totalThreshold = 0)

Drops all null fields in a `DynamicFrame` whose type is `NullType`. These are fields with missing or null values in every record in the `DynamicFrame` dataset.

- `frame` – The `DynamicFrame` to drop null fields in (required).
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string associated with errors in the transformation (optional).
- `stageThreshold` – The maximum number of errors that can occur in the transformation before it errors out (optional). The default is zero.
- `totalThreshold` – The maximum number of errors that can occur overall before processing errors out (optional). The default is zero.

Returns a new `DynamicFrame` with no null fields.

apply(cls, *args, **kwargs)

- `cls` – cls

name(cls)

- `cls` – cls

describeArgs(cls)

- `cls` – cls

describeReturn(cls)

- `cls` – cls

describeTransform(cls)

- `cls` – cls

describeErrors(cls)

- `cls` – cls

describe(cls)

- `cls` – cls

ErrorsAsDynamicFrame class

Returns a `DynamicFrame` that contains nested records for errors that occurred while the source `DynamicFrame` was created.

Example

We recommend that you use the `DynamicFrame.errorsAsDynamicFrame()` method to retrieve and view error records. To view a code example, see Example: Use errorsAsDynamicFrame to view error records.

Methods

- __call__
- apply
- name
- describeArgs
- describeReturn
- describeTransform
- describeErrors
- Describe

__call__(frame)

Returns a `DynamicFrame` that contains nested error records that relate to the source `DynamicFrame`.

- `frame` – The source `DynamicFrame` (required).

apply(cls, *args, **kwargs)

- `cls` – cls

name(cls)

- `cls` – cls

describeArgs(cls)

- `cls` – cls

describeReturn(cls)

- `cls` – cls

describeTransform(cls)

- `cls` – cls

describeErrors(cls)

- `cls` – cls

describe(cls)

- `cls` – cls

EvaluateDataQuality class

Evaluates a data quality ruleset against a `DynamicFrame` and returns a new `DynamicFrame` with results of the evaluation.

Example

The following example code demonstrates how to evaluate data quality for a `DynamicFrame` and then view the data quality results.

```
from awsglue.transforms import *
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsgluedq.transforms import EvaluateDataQuality

#Create Glue context
sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)

# Define DynamicFrame
legislatorsAreas = glueContext.create_dynamic_frame.from_catalog(
    database="legislators", table_name="areas_json")

# Create data quality ruleset
ruleset = """Rules = [ColumnExists "id", IsComplete "id"]"""

# Evaluate data quality
dqResults = EvaluateDataQuality.apply(
    frame=legislatorsAreas,
    ruleset=ruleset,
    publishing_options={
        "dataQualityEvaluationContext": "legislatorsAreas",
        "enableDataQualityCloudWatchMetrics": True,
        "enableDataQualityResultsPublishing": True,
        "resultsS3Prefix": "amzn-s3-demo-bucket1",
    },
)


# Inspect data quality results
dqResults.printSchema()
dqResults.toDF().show()
```

Output

```
root
|-- Rule: string
|-- Outcome: string
|-- FailureReason: string
|-- EvaluatedMetrics: map
|    |-- keyType: string
|    |-- valueType: double


+-----------------------+-------+-------------+---------------------------------------+
|Rule                   |Outcome|FailureReason|EvaluatedMetrics                       |
+-----------------------+-------+-------------+---------------------------------------+
|ColumnExists "id"      |Passed |null         |{}                                     |
|IsComplete "id"        |Passed |null         |{Column.first_name.Completeness -> 1.0}|
+-----------------------+-------+-------------+---------------------------------------+
```

Methods

- __call__
- apply
- name
- describeArgs
- describeReturn
- describeTransform
- describeErrors
- describe

__call__(frame, ruleset, publishing_options = {})

- `frame` – The `DynamicFrame` that you want evaluate the data quality of.
- `ruleset` – A Data Quality Definition Language (DQDL) ruleset in string format. To learn more about DQDL, see the Data Quality Definition Language (DQDL) reference guide.
- `publishing_options` – A dictionary that specifies the following options for publishing evaluation results and metrics:

- `dataQualityEvaluationContext` – A string that specifies the namespace under which AWS Glue should publish Amazon CloudWatch metrics and the data quality results. The aggregated metrics appear in CloudWatch, while the full results appear in the AWS Glue Studio interface.
  - Required: No
  - Default value: `default_context`
- `enableDataQualityCloudWatchMetrics` – Specifies whether the results of the data quality evaluation should be published to CloudWatch. You specify a namespace for the metrics using the `dataQualityEvaluationContext` option.
  - Required: No
  - Default value: False
- `enableDataQualityResultsPublishing` – Specifies whether the data quality results should be visible on the **Data Quality** tab in the AWS Glue Studio interface.
  - Required: No
  - Default value: True
- `resultsS3Prefix` – Specifies the Amazon S3 location where AWS Glue can write the data quality evaluation results.
  - Required: No
  - Default value: "" (empty string)

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

FillMissingValues class

The `FillMissingValues` class locates null values and empty strings in a specified `DynamicFrame` and uses machine learning methods, such as linear regression and random forest, to predict the missing values. The ETL job uses the values in the input dataset to train the machine learning model, which then predicts what the missing values should be.

> **Tip**
>
> If you use incremental data sets, then each incremental set is used as the training data for the machine learning model, so the results might not be as accurate.

To import:

```
from awsglueml.transforms import FillMissingValues
```

Methods

- Apply

apply(frame, missing_values_column, output_column ="", transformation_ctx ="", info ="", stageThreshold = 0, totalThreshold = 0)

Fills a dynamic frame's missing values in a specified column and returns a new frame with estimates in a new column. For rows without missing values, the specified column's value is duplicated to the new column.

- `frame` – The `DynamicFrame` in which to fill missing values. Required.
- `missing_values_column` – The column containing missing values (`null` values and empty strings). Required.

- `output_column` – The name of the new column that will contain estimated values for all rows whose value was missing. Optional; the default is the name of `missing_values_column` suffixed by `"_filled"`.
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string associated with errors in the transformation (optional).
- `stageThreshold` – The maximum number of errors that can occur in the transformation before it errors out (optional; the default is zero).
- `totalThreshold` – The maximum number of errors that can occur overall before processing errors out (optional; the default is zero).

Returns a new `DynamicFrame` with one additional column that contains estimations for rows with missing values and the present value for other rows.

Filter class

Builds a new `DynamicFrame` that contains records from the input `DynamicFrame` that satisfy a specified predicate function.

Example

We recommend that you use the `DynamicFrame.filter()` method to filter records in a `DynamicFrame`. To view a code example, see Example: Use filter to get a filtered selection of fields.

Methods

- __call__
- apply
- name
- describeArgs
- describeReturn
- describeTransform
- describeErrors
- describe

__call__(frame, f, transformation_ctx="", info="", stageThreshold=0, totalThreshold=0))

Returns a new `DynamicFrame` that is built by selecting records from the input `DynamicFrame` that satisfy a specified predicate function.

- `frame` – The source `DynamicFrame` to apply the specified filter function to (required).
- `f` – The predicate function to apply to each `DynamicRecord` in the `DynamicFrame`. The function must take a `DynamicRecord` as its argument and return True if the `DynamicRecord` meets the filter requirements, or False if it doesn't (required).

  A `DynamicRecord` represents a logical record in a `DynamicFrame`. It's similar to a row in a Spark `DataFrame`, except that it is self-describing and can be used for data that doesn't conform to a fixed schema.
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string that is associated with errors in the transformation (optional).
- `stageThreshold` – The maximum number of errors that can occur in the transformation before it errors out (optional). The default is zero.
- `totalThreshold` – The maximum number of errors that can occur overall before processing errors out (optional). The default is zero.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

FindIncrementalMatches class

Identifies matching records in the existing and incremental `DynamicFrame` and creates a new `DynamicFrame` with a unique identifier assigned to each group of matching records.

To import:

```
from awsglueml.transforms import FindIncrementalMatches
```

Methods

- Apply

apply(existingFrame, incrementalFrame, transformId, transformation_ctx = "", info = "", stageThreshold = 0, totalThreshold = 0, enforcedMatches = none, computeMatchConfidenceScores = 0)

Identifies matching records in the input `DynamicFrame` and creates a new `DynamicFrame` with a unique identifier assigned to each group of matching records.

- `existingFrame` – The existing and pre-matched `DynamicFrame` to apply the FindIncrementalMatches transform. Required.
- `incrementalFrame` – The incremental `DynamicFrame` to apply the FindIncrementalMatches transform to match against the `existingFrame`. Required.
- `transformId` – The unique ID associated with the FindIncrementalMatches transform to apply on records in the `DynamicFrames`. Required.
- `transformation_ctx` – A unique string that is used to identify stats/state information. Optional.
- `info` – A string to be associated with errors in the transformation. Optional.
- `stageThreshold` – The maximum number of errors that can occur in the transformation before it errors out. Optional. The default is zero.

- `totalThreshold` – The maximum number of errors that can occur overall before processing errors out. Optional. The default is zero.
- `enforcedMatches` – The `DynamicFrame` used to enforce matches. Optional. The default is None.
- `computeMatchConfidenceScores` – A Boolean value indicating whether to compute a confidence score for each group of matching records. Optional. The default is false.

Returns a new `DynamicFrame` with a unique identifier assigned to each group of matching records.

FindMatches class

Identifies matching records in the input `DynamicFrame` and creates a new `DynamicFrame` with a unique identifier assigned to each group of matching records.

To import:

```
from awsglueml.transforms import FindMatches
```

Methods

- Apply

apply(frame, transformId, transformation_ctx = "", info = "", stageThreshold = 0, totalThreshold = 0, enforcedMatches = none, computeMatchConfidenceScores = 0)

Identifies matching records in the input `DynamicFrame` and creates a new `DynamicFrame` with a unique identifier assigned to each group of matching records.

- `frame` – The `DynamicFrame` to apply the FindMatches transform. Required.
- `transformId` – The unique ID associated with the FindMatches transform to apply on records in the `DynamicFrame`. Required.
- `transformation_ctx` – A unique string that is used to identify stats/state information. Optional.
- `info` – A string to be associated with errors in the transformation. Optional.
- `stageThreshold` – The maximum number of errors that can occur in the transformation before it errors out. Optional. The default is zero.

- `totalThreshold` – The maximum number of errors that can occur overall before processing errors out. Optional. The default is zero.
- `enforcedMatches` – The `DynamicFrame` used to enforce matches. Optional. The default is None.
- `computeMatchConfidenceScores` – A Boolean value indicating whether to compute a confidence score for each group of matching records. Optional. The default is false.

Returns a new `DynamicFrame` with a unique identifier assigned to each group of matching records.

FlatMap class

Applies a transform to each `DynamicFrame` in a collection. Results are not flattened into a single `DynamicFrame`, but preserved as a collection.

Examples for FlatMap

The following example snippet demonstrates how to use the `ResolveChoice` transform on a collection of dynamic frames when applied to a `FlatMap`. The data used for input is in the JSON located at the placeholder Amazon S3 address `s3://bucket/path-for-data/sample.json` and contains the following data.

Example JSON data

```
[{
    "firstname": "Arnav",
    "lastname": "Desai",
    "address": {
        "street": "6 Anyroad Avenue",
        "city": "London",
        "state": "England",
        "country": "UK"
    },
    "phone": 17235550101,
    "affiliations": [
        "General Anonymous Example Products",
        "Example Independent Research",
        "Government Department of Examples"
    ]
},
{
    "firstname": "Mary",
    "lastname": "Major",
    "address": {
        "street": "7821 Spot Place",
        "city": "Centerville",
        "state": "OK",
        "country": "US"
    },
    "phone": 19185550023,
    "affiliations": [
        "Example Dot Com",
        "Example Independent Research",
        "Example.io"
    ]
},
{
    "firstname": "Paulo",
    "lastname": "Santos",
    "address": {
        "street": "123 Maple Street",
        "city": "London",
        "state": "Ontario",
        "country": "CA"
    },
    "phone": 12175550181,
    "affiliations": [
        "General Anonymous Example Products",
        "Example Dot Com"
    ]
}]
```

Example Apply ResolveChoice to a DynamicFrameCollection and show output.

```
#Read DynamicFrame
datasource = glueContext.create_dynamic_frame_from_options("s3", connection_options =
 {"paths":["s3://bucket/path/to/file/mysamplejson.json"]}, format="json")
datasource.printSchema()
datasource.show()

## Split to create a DynamicFrameCollection
split_frame=datasource.split_fields(["firstname","lastname","address"],"personal_info","busines
split_frame.keys()
print("---")
```

```
## Use FlatMap to run ResolveChoice
kwargs = {"choice": "cast:string"}
flat = FlatMap.apply(split_frame, ResolveChoice, frame_name="frame",
 transformation_ctx='tcx', **kwargs)
flat.keys()

##Select one of the DynamicFrames
personal_info = flat.select("personal_info")
personal_info.printSchema()
personal_info.show()
print("---")

business_info = flat.select("business_info")
business_info.printSchema()
business_info.show()
```

> **Important**
>
> When calling `FlatMap.apply`, the `frame_name` parameter **must** be `"frame"`. No other value is currently accepted.

Example output

```
root
|-- firstname: string
|-- lastname: string
|-- address: struct
|    |-- street: string
|    |-- city: string
|    |-- state: string
|    |-- country: string
|-- phone: long
|-- affiliations: array
|    |-- element: string
---
{
    "firstname": "Mary",
    "lastname": "Major",
    "address": {
        "street": "7821 Spot Place",
        "city": "Centerville",
        "state": "OK",
        "country": "US"
    },
    "phone": 19185550023,
    "affiliations": [
        "Example Dot Com",
        "Example Independent Research",
        "Example.io"
    ]
}

{
    "firstname": "Paulo",
    "lastname": "Santos",
    "address": {
        "street": "123 Maple Street",
        "city": "London",
        "state": "Ontario",
        "country": "CA"
    },
    "phone": 12175550181,
    "affiliations": [
        "General Anonymous Example Products",
        "Example Dot Com"
    ]
}
---
root
|-- firstname: string
|-- lastname: string
|-- address: struct
|    |-- street: string
|    |-- city: string
|    |-- state: string
|    |-- country: string

{
    "firstname": "Mary",
    "lastname": "Major",
    "address": {
        "street": "7821 Spot Place",
        "city": "Centerville",
        "state": "OK",
        "country": "US"
    }
}

{
    "firstname": "Paulo",
    "lastname": "Santos",
    "address": {
        "street": "123 Maple Street",
        "city": "London",
        "state": "Ontario",
        "country": "CA"
    }
}
---
root
|-- phone: long
|-- affiliations: array
|    |-- element: string

{
    "phone": 19185550023,
    "affiliations": [
        "Example Dot Com",
        "Example Independent Research",
        "Example.io"
    ]
}

{
    "phone": 12175550181,
    "affiliations": [
        "General Anonymous Example Products",
        "Example Dot Com"
    ]
}
```

Methods

- __call__
- Apply
- Name
- describeArgs

- describeReturn
- describeTransform
- describeErrors
- Describe

__call__(dfc, BaseTransform, frame_name, transformation_ctx = "", **base_kwargs)

Applies a transform to each `DynamicFrame` in a collection and flattens the results.

- `dfc` – The `DynamicFrameCollection` over which to flatmap (required).
- `BaseTransform` – A transform derived from `GlueTransform` to apply to each member of the collection (required).
- `frame_name` – The argument name to pass the elements of the collection to (required).
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `base_kwargs` – Arguments to pass to the base transform (required).

Returns a new `DynamicFrameCollection` created by applying the transform to each `DynamicFrame` in the source `DynamicFrameCollection`.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

Join class

Performs an equality join on two `DynamicFrames`.

Example

We recommend that you use the `DynamicFrame.join()` method to join `DynamicFrames`. To view a code example, see Example: Use join to combine `DynamicFrames`.

Methods

- __call__
- apply
- name
- describeArgs
- describeReturn
- describeTransform
- describeErrors
- Describe

__call__(frame1, frame2, keys1, keys2, transformation_ctx = "")

Performs an equality join on two `DynamicFrames`.

- `frame1` – The first `DynamicFrame` to join (required).
- `frame2` – The second `DynamicFrame` to join (required).
- `keys1` – The keys to join on for the first frame (required).
- `keys2` – The keys to join on for the second frame (required).
- `transformation_ctx` – A unique string that is used to identify state information (optional).

Returns a new `DynamicFrame` that is created by joining the two `DynamicFrames`.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

Map class

Builds a new `DynamicFrame` by applying a function to all records in the input `DynamicFrame`.

Example

We recommend that you use the `DynamicFrame.map()` method to apply a function to all records in a `DynamicFrame`. To view a code example, see Example: Use map to apply a function to every record in a `DynamicFrame`.

Methods

- __call__
- apply
- name

- describeArgs
- describeReturn
- describeTransform
- describeErrors
- Describe

__call__(frame, f, transformation_ctx="", info="", stageThreshold=0, totalThreshold=0)

Returns a new `DynamicFrame` that results from applying the specified function to all `DynamicRecords` in the original `DynamicFrame`.

- `frame` – The original `DynamicFrame` to apply the mapping function to (required).
- `f` – The function to apply to all `DynamicRecords` in the `DynamicFrame`. The function must take a `DynamicRecord` as an argument and return a new `DynamicRecord` that is produced by the mapping (required).

  A `DynamicRecord` represents a logical record in a `DynamicFrame`. It's similar to a row in an Apache Spark `DataFrame`, except that it is self-describing and can be used for data that doesn't conform to a fixed schema.
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string associated with errors in the transformation (optional).
- `stageThreshold` – The maximum number of errors that can occur in the transformation before it errors out (optional). The default is zero.
- `totalThreshold` – The maximum number of errors that can occur overall before processing errors out (optional). The default is zero.

Returns a new `DynamicFrame` that results from applying the specified function to all `DynamicRecords` in the original `DynamicFrame`.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

MapToCollection class

Applies a transform to each `DynamicFrame` in the specified `DynamicFrameCollection`.

Methods

- __call__
- Apply
- Name
- describeArgs
- describeReturn
- describeTransform
- describeErrors
- Describe

__call__(dfc, BaseTransform, frame_name, transformation_ctx = "", **base_kwargs)

Applies a transform function to each `DynamicFrame` in the specified `DynamicFrameCollection`.

- `dfc` – The `DynamicFrameCollection` over which to apply the transform function (required).
- `callable` – A callable transform function to apply to each member of the collection (required).

- `transformation_ctx` – A unique string that is used to identify state information (optional).

Returns a new `DynamicFrameCollection` created by applying the transform to each `DynamicFrame` in the source `DynamicFrameCollection`.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

Relationalize class

Flattens a nested schema in a `DynamicFrame` and pivots out array columns from the flattened frame.

Example

We recommend that you use the `DynamicFrame.relationalize()` method to relationalize a `DynamicFrame`. To view a code example, see Example: Use relationalize to flatten a nested schema in a `DynamicFrame`.

Methods

- __call__
- apply
- name
- describeArgs
- describeReturn
- describeTransform
- describeErrors
- describe

__call__(frame, staging_path=None, name='roottable', options=None, transformation_ctx = "", info = "", stageThreshold = 0, totalThreshold = 0)

Relationalizes a `DynamicFrame` and produces a list of frames that are generated by unnesting nested columns and pivoting array columns. You can join a pivoted array column to the root table by using the join key that is generated in the unnest phase.

- `frame` – The `DynamicFrame` to relationalize (required).
- `staging_path` – The path where the method can store partitions of pivoted tables in CSV format (optional). Pivoted tables are read back from this path.
- `name` – The name of the root table (optional).
- `options` – A dictionary of optional parameters. Currently unused.
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string associated with errors in the transformation (optional).
- `stageThreshold` – The maximum number of errors that can occur in the transformation before it errors out (optional). The default is zero.
- `totalThreshold` – The maximum number of errors that can occur overall before processing errors out (optional). The default is zero.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

RenameField class

Renames a node within a `DynamicFrame`.

Example

We recommend that you use the `DynamicFrame.rename_field()` method to rename a field in a `DynamicFrame`. To view a code example, see Example: Use rename_field to rename fields in a `DynamicFrame`.

Methods

- __call__
- apply
- name
- describeArgs
- describeReturn
- describeTransform
- describeErrors

- describe

__call__(frame, old_name, new_name, transformation_ctx = "", info = "", stageThreshold = 0, totalThreshold = 0)

Renames a node within a `DynamicFrame`.

- `frame` – The `DynamicFrame` in which to rename a node (required).
- `old_name` – The full path to the node to rename (required).

  If the old name has dots in it, RenameField will not work unless you place backticks around it (``` `` ```). For example, to replace `this.old.name` with `thisNewName`, you would call RenameField as follows:

  ```
  newDyF = RenameField(oldDyF, "`this.old.name`", "thisNewName")
  ```
- `new_name` – The new name, including full path (required).
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string associated with errors in the transformation (optional).
- `stageThreshold` – The maximum number of errors that can occur in the transformation before it errors out (optional). The default is zero.
- `totalThreshold` – The maximum number of errors that can occur overall before processing errors out (optional). The default is zero.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

ResolveChoice class

Resolves a choice type within a `DynamicFrame`.

Example

We recommend that you use the `DynamicFrame.resolveChoice()` method to handle fields that contain multiple types in a `DynamicFrame`. To view a code example, see Example: Use resolveChoice to handle a column that contains multiple types.

Methods

- __call__
- apply
- name
- describeArgs
- describeReturn
- describeTransform
- describeErrors
- describe

__call__(frame, specs = none, choice = "", transformation_ctx = "", info = "", stageThreshold = 0, totalThreshold = 0)

Provides information for resolving ambiguous types within a `DynamicFrame`. It returns the resulting `DynamicFrame`.

- `frame` – The `DynamicFrame` in which to resolve the choice type (required).
- `specs` – A list of specific ambiguities to resolve, each in the form of a tuple: `(path, action)`. The `path` value identifies a specific ambiguous element, and the `action` value identifies the corresponding resolution.

  You can only use one of the `spec` and `choice` parameters. If the `spec` parameter is not `None`, then the `choice` parameter must be an empty string. Conversely, if the `choice` is not an empty string, then the `spec` parameter must be `None`. If neither parameter is provided, AWS Glue tries to parse the schema and use it to resolve ambiguities.

  You can specify one of the following resolution strategies in the `action` portion of a `specs` tuple:

  - `cast` – Allows you to specify a type to cast to (for example, `cast:int`).
  - `make_cols` – Resolves a potential ambiguity by flattening the data. For example, if `columnA` could be an `int` or a `string`, the resolution is to produce two columns named `columnA_int` and `columnA_string` in the resulting `DynamicFrame`.
  - `make_struct` – Resolves a potential ambiguity by using a struct to represent the data. For example, if data in a column could be an `int` or a `string`, using the `make_struct` action produces a column of structures in the resulting `DynamicFrame` with each containing both an `int` and a `string`.
  - `project` – Resolves a potential ambiguity by retaining only values of a specified type in the resulting `DynamicFrame`. For example, if data in a `ChoiceType` column could be an `int` or a `string`, specifying a `project:string` action drops values from the resulting `DynamicFrame` that are not type `string`.

  If the `path` identifies an array, place empty square brackets after the name of the array to avoid ambiguity. For example, suppose you are working with data structured as follows:

  ```
  "myList": [
    { "price": 100.00 },
    { "price": "$100.00" }
  ]
  ```

  You can select the numeric rather than the string version of the price by setting the `path` to `"myList[].price"`, and setting the `action` to `"cast:double"`.
- `choice` – The default resolution action if the `specs` parameter is `None`. If the `specs` parameter is not `None`, then this must not be set to anything but an empty string.

In addition to the `specs` actions previously described, this argument also supports the following action:

  - `MATCH_CATALOG` – Attempts to cast each `ChoiceType` to the corresponding type in the specified Data Catalog table.
- `database` – The AWS Glue Data Catalog database to use with the `MATCH_CATALOG` choice (required for `MATCH_CATALOG`).
- `table_name` – The AWS Glue Data Catalog table name to use with the `MATCH_CATALOG` action (required for `MATCH_CATALOG`).
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string associated with errors in the transformation (optional).
- `stageThreshold` – The maximum number of errors that can occur in the transformation before it errors out (optional). The default is zero.
- `totalThreshold` – The maximum number of errors that can occur overall before processing errors out (optional). The default is zero.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

SelectFields class

The `SelectFields` class creates a new `DynamicFrame` from an existing `DynamicFrame`, and keeps only the fields that you specify. `SelectFields` provides similar functionality to a SQL `SELECT` statement.

Example

We recommend that you use the `DynamicFrame.select_fields()` method to select fields from a `DynamicFrame`. To view a code example, see Example: Use select_fields to create a new `DynamicFrame` with chosen fields.

Methods

- __call__
- apply
- name
- describeArgs
- describeReturn
- describeTransform
- describeErrors
- Describe

__call__(frame, paths, transformation_ctx = "", info = "", stageThreshold = 0, totalThreshold = 0)

Gets fields (nodes) in a `DynamicFrame`.

- `frame` – The `DynamicFrame` to select fields in (required).
- `paths` – A list of full paths to the fields to select (required).
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string that is associated with errors in the transformation (optional).
- `stageThreshold` – The maximum number of errors that can occur in the transformation before it errors out (optional). The default is zero.

- `totalThreshold` – The maximum number of errors that can occur overall before processing errors out (optional). The default is zero.

Returns a new `DynamicFrame` that contains only the specified fields.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

SelectFromCollection class

Selects one `DynamicFrame` in a `DynamicFrameCollection`.

Example

This example uses `SelectFromCollection` to select a `DynamicFrame` from a `DynamicFrameCollection`.

Example dataset

The example selects two `DynamicFrames` from a `DynamicFrameCollection` called `split_rows_collection`. The following is the list of keys in `split_rows_collection`.

```
dict_keys(['high', 'low'])
```

Example code

```
# Example: Use SelectFromCollection to select
# DynamicFrames from a DynamicFrameCollection

from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.transforms import SelectFromCollection

# Create GlueContext
sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)

# Select frames and inspect entries
frame_low = SelectFromCollection.apply(dfc=split_rows_collection, key="low")
frame_low.toDF().show()

frame_high = SelectFromCollection.apply(dfc=split_rows_collection, key="high")
frame_high.toDF().show()
```

Output

```
+---+-----+------------------------+-------------------------+
| id|index|contact_details.val.type|contact_details.val.value|
+---+-----+------------------------+-------------------------+
|  1|    0|                     fax|             202-225-3307|
|  1|    1|                   phone|             202-225-5731|
|  2|    0|                     fax|             202-225-3307|
|  2|    1|                   phone|             202-225-5731|
|  3|    0|                     fax|             202-225-3307|
|  3|    1|                   phone|             202-225-5731|
|  4|    0|                     fax|             202-225-3307|
|  4|    1|                   phone|             202-225-5731|
|  5|    0|                     fax|             202-225-3307|
|  5|    1|                   phone|             202-225-5731|
|  6|    0|                     fax|             202-225-3307|
|  6|    1|                   phone|             202-225-5731|
|  7|    0|                     fax|             202-225-3307|
|  7|    1|                   phone|             202-225-5731|
|  8|    0|                     fax|             202-225-3307|
|  8|    1|                   phone|             202-225-5731|
|  9|    0|                     fax|             202-225-3307|
|  9|    1|                   phone|             202-225-5731|
| 10|    0|                     fax|             202-225-6328|
| 10|    1|                   phone|             202-225-4576|
+---+-----+------------------------+-------------------------+
only showing top 20 rows

+---+-----+------------------------+-------------------------+
| id|index|contact_details.val.type|contact_details.val.value|
+---+-----+------------------------+-------------------------+
| 11|    0|                     fax|             202-225-6328|
| 11|    1|                   phone|             202-225-4576|
| 11|    2|                 twitter|           RepTrentFranks|
| 12|    0|                     fax|             202-225-6328|
| 12|    1|                   phone|             202-225-4576|
| 12|    2|                 twitter|           RepTrentFranks|
| 13|    0|                     fax|             202-225-6328|
| 13|    1|                   phone|             202-225-4576|
| 13|    2|                 twitter|           RepTrentFranks|
| 14|    0|                     fax|             202-225-6328|
| 14|    1|                   phone|             202-225-4576|
| 14|    2|                 twitter|           RepTrentFranks|
| 15|    0|                     fax|             202-225-6328|
| 15|    1|                   phone|             202-225-4576|
| 15|    2|                 twitter|           RepTrentFranks|
| 16|    0|                     fax|             202-225-6328|
| 16|    1|                   phone|             202-225-4576|
| 16|    2|                 twitter|           RepTrentFranks|
| 17|    0|                     fax|             202-225-6328|
| 17|    1|                   phone|             202-225-4576|
+---+-----+------------------------+-------------------------+
only showing top 20 rows
```

Methods

- __call__
- apply
- name

- describeArgs
- describeReturn
- describeTransform
- describeErrors
- describe

__call__(dfc, key, transformation_ctx = "")

Gets one `DynamicFrame` from a `DynamicFrameCollection`.

- `dfc` – The `DynamicFrameCollection` that the `DynamicFrame` should be selected from (required).
- `key` – The key of the `DynamicFrame` to select (required).
- `transformation_ctx` – A unique string that is used to identify state information (optional).

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

Simplify_ddb_json class

Simplifies nested columns in a `DynamicFrame` that are specifically in the DynamoDB JSON structure, and returns a new simplified `DynamicFrame`.

Example

We recommend that you use the `DynamicFrame.simplify_ddb_json()` method to simplify nested columns in a `DynamicFrame` that are specifically in the DynamoDB JSON structure. To view a code example, see Example: Use simplify_ddb_json to invoke a DynamoDB JSON simplify.

Spigot class

Writes sample records to a specified destination to help you verify the transformations performed by your AWS Glue job.

Example

We recommend that you use the `DynamicFrame.spigot()` method to write a subset of records from a `DynamicFrame` to a specified destination. To view a code example, see Example: Use spigot to write sample fields from a `DynamicFrame` to Amazon S3.

Methods

- __call__
- apply
- name
- describeArgs
- describeReturn
- describeTransform
- describeErrors
- describe

__call__(frame, path, options, transformation_ctx = "")

Writes sample records to a specified destination during a transformation.

- `frame` – The `DynamicFrame` to spigot (required).
- `path` – The path of the destination to write to (required).
- `options` – JSON key-value pairs that specify options (optional). The `"topk"` option specifies that the first *k* records should be written. The `"prob"` option specifies the probability (as a decimal) of picking any given record. You use this in selecting records to write.
- `transformation_ctx` – A unique string that is used to identify state information (optional).

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply

name(cls)

Inherited from `GlueTransform` name

describeArgs(cls)

Inherited from `GlueTransform` describeArgs

describeReturn(cls)

Inherited from `GlueTransform` describeReturn

describeTransform(cls)

Inherited from `GlueTransform` describeTransform

describeErrors(cls)

Inherited from `GlueTransform` describeErrors

describe(cls)

Inherited from `GlueTransform` describe

SplitFields class

Splits a `DynamicFrame` into two new ones, by specified fields.

Example

We recommend that you use the `DynamicFrame.split_fields()` method to split fields in a `DynamicFrame`. To view a code example, see Example: Use split_fields to split selected fields into a separate `DynamicFrame`.

Methods

- __call__
- apply
- name
- describeArgs
- describeReturn
- describeTransform
- describeErrors
- describe

__call__(frame, paths, name1 = none, name2 = none, transformation_ctx = "", info = "", stageThreshold = 0, totalThreshold = 0)

Splits one or more fields in a `DynamicFrame` off into a new `DynamicFrame`, and creates another new `DynamicFrame` that contains the fields that remain.

- `frame` – The source `DynamicFrame` to split into two new ones (required).
- `paths` – A list of full paths to the fields to be split (required).
- `name1` – The name to assign to the `DynamicFrame` that will contain the fields to be split off (optional). If no name is supplied, the name of the source frame is used with "1" appended.
- `name2` – The name to assign to the `DynamicFrame` that will contain the fields that remain after the specified fields are split off (optional). If no name is provided, the name of the source frame is used with "2" appended.
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string associated with errors in the transformation (optional).
- `stageThreshold` – The maximum number of errors that can occur in the transformation before it errors out (optional). The default is zero.

- `totalThreshold` – The maximum number of errors that can occur overall before processing errors out (optional). The default is zero.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

SplitRows class

Creates a `DynamicFrameCollection` that contains two `DynamicFrames`. One `DynamicFrame` contains only the specified rows to be split, and the other contains all remaining rows.

Example

We recommend that you use the `DynamicFrame.split_rows()` method to split rows in a `DynamicFrame`. To view a code example, see Example: Use split_rows to split rows in a `DynamicFrame`.

Methods

- __call__

- apply
- name
- describeArgs
- describeReturn
- describeTransform
- describeErrors
- describe

__call__(frame, comparison_dict, name1="frame1", name2="frame2", transformation_ctx = "", info = none, stageThreshold = 0, totalThreshold = 0)

Splits one or more rows in a `DynamicFrame` off into a new `DynamicFrame`.

- `frame` – The source `DynamicFrame` to split into two new ones (required).
- `comparison_dict` – A dictionary where the key is the full path to a column, and the value is another dictionary for mapping comparators to values that the column values are compared to. For example, `{"age": {">": 10, "<": 20}}` splits rows where the value of "age" is between 10 and 20, exclusive, from rows where "age" is outside that range (required).
- `name1` – The name to assign to the `DynamicFrame` that will contain the rows to be split off (optional).
- `name2` – The name to assign to the `DynamicFrame` that will contain the rows that remain after the specified rows are split off (optional).
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string associated with errors in the transformation (optional).
- `stageThreshold` – The maximum number of errors that can occur in the transformation before it errors out (optional). The default is zero.
- `totalThreshold` – The maximum number of errors that can occur overall before processing errors out (optional). The default is zero.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

Unbox class

Unboxes (reformats) a string field in a `DynamicFrame`.

Example

We recommend that you use the `DynamicFrame.unbox()` method to unbox a field in a `DynamicFrame`. To view a code example, see Example: Use unbox to unbox a string field into a struct.

Methods

- __call__
- apply
- name
- describeArgs
- describeReturn

- describeTransform
- describeErrors
- describe

__call__(frame, path, format, transformation_ctx = "", info="", stageThreshold=0, totalThreshold=0, **options)

Unboxes a string field in a `DynamicFrame`.

- `frame` – The `DynamicFrame` in which to unbox a field. (required).
- `path` – The full path to the `StringNode` to unbox (required).
- `format` – A format specification (optional). This is used for an Amazon S3 or AWS Glue connection that supports multiple formats. For the formats that are supported, see Data format options for inputs and outputs in AWS Glue for Spark.
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string associated with errors in the transformation (optional).
- `stageThreshold` – The maximum number of errors that can occur in the transformation before it errors out (optional). The default is zero.
- `totalThreshold` – The maximum number of errors that can occur overall before processing errors out (optional). The default is zero.
- `separator` – A separator token (optional).
- `escaper` – An escape token (optional).
- `skipFirst` – `True` if the first line of data should be skipped, or `False` if it should not be skipped (optional).
- withSchema – A string that contains a schema for the data to be unboxed (optional). This should always be created using `StructType.json`.
- `withHeader` – `True` if the data being unpacked includes a header, or `False` if not (optional).

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

UnnestFrame class

Unnests a `DynamicFrame`, flattens nested objects to top-level elements, and generates join keys for array objects.

Example

We recommend that you use the `DynamicFrame.unnest()` method to flatten nested structures in a `DynamicFrame`. To view a code example, see Example: Use unnest to turn nested fields into top-level fields.

Methods

- __call__
- apply
- name
- describeArgs
- describeReturn
- describeTransform

- describeErrors
- describe

__call__(frame, transformation_ctx = "", info="", stageThreshold=0, totalThreshold=0)

Unnests a `DynamicFrame`, flattens nested objects to top-level elements, and generates join keys for array objects.

- `frame` – The `DynamicFrame` to unnest (required).
- `transformation_ctx` – A unique string that is used to identify state information (optional).
- `info` – A string associated with errors in the transformation (optional).
- `stageThreshold` – The maximum number of errors that can occur in the transformation before it errors out (optional). The default is zero.
- `totalThreshold` – The maximum number of errors that can occur overall before processing errors out (optional). The default is zero.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

FlagDuplicatesInColumn class

The `FlagDuplicatesInColumn` transform returns a new column with a specified value in each row that indicates whether the value in the row's source column matches a value in an earlier row of the source column. When matches are found, they are flagged as duplicates. The initial occurrence is not flagged, because it doesn't match an earlier row.

Example

```
from pyspark.context import SparkContext
from pyspark.sql import SparkSession
from awsgluedi.transforms import *

sc = SparkContext()
spark = SparkSession(sc)

datasource1 = spark.read.json("s3://${BUCKET}/json/zips/raw/data")

try:
    df_output = column.FlagDuplicatesInColumn.apply(
        data_frame=datasource1,
        spark_context=sc,
        source_column="city",
        target_column="flag_col",
        true_string="True",
        false_string="False"
    )
except:
    print("Unexpected Error happened ")
    raise
```

Output

The `FlagDuplicatesInColumn` transformation will add a new column `flag_col` to the `df_output` DataFrame. This column will contain a string value indicating whether the corresponding row has a duplicate value in the `city` column or not. If a row has a duplicate `city` value, the `flag_col` will contain the `true_string` value "True". If a row has a unique `city` value, the `flag_col` will contain the `false_string` value "False".

The resulting `df_output` DataFrame will contain all columns from the original `datasource1` DataFrame, plus the additional `flag_col` column indicating duplicate `city` values.

Methods

- __call__
- apply
- name
- describeArgs
- describeReturn
- describeTransform
- describeErrors
- describe

__call__(spark_context, data_frame, source_column, target_column, true_string=DEFAULT_TRUE_STRING, false_string=DEFAULT_FALSE_STRING)

The `FlagDuplicatesInColumn` transform returns a new column with a specified value in each row that indicates whether the value in the row's source column matches a value in an earlier row of the source column. When matches are found, they are flagged as duplicates. The initial occurrence is not flagged, because it doesn't match an earlier row.

- `source_column` – Name of the source column.
- `target_column` – Name of the target column.
- `true_string` – String to be inserted in the target column when a source column value duplicates an earlier value in that column.
- `false_string` – String to be inserted in the target column when a source column value is distinct from earlier values in that column.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

FormatPhoneNumber class

The `FormatPhoneNumber` transform returns a column in which a phone number string is converted into a formatted value.

Example

```
from pyspark.context import SparkContext
from pyspark.sql import SparkSession
from awsgluedi.transforms import *

sc = SparkContext()
spark = SparkSession(sc)

input_df = spark.createDataFrame(
    [
        ("408-341-5669",),
        ("4083415669",)
    ],
    ["phone"],
)
```

```
try:
    df_output = column_formatting.FormatPhoneNumber.apply(
        data_frame=input_df,
        spark_context=sc,
        source_column="phone",
        default_region="US"
    )
    df_output.show()
except:
    print("Unexpected Error happened ")
    raise
```

Output

The output will be:

```
```
+---------------+
| phone|
+---------------+
|(408) 341-5669|
|(408) 341-5669|
+---------------+
```
```

The `FormatPhoneNumber` transformation takes the `source_column` as `"phone"` and the `default_region` as `"US"`.

The transformation successfully formats both phone numbers, regardless of their initial format, to the standard US format `(408) 341-5669`.

Methods

- __call__
- apply
- name
- describeArgs
- describeReturn
- describeTransform

- describeErrors
- describe

__call__(spark_context, data_frame, source_column, phone_number_format=None, default_region=None, default_region_column=None)

The `FormatPhoneNumber` transform returns a column in which a phone number string is converted into a formatted value.

- `source_column` – The name of an existing column.
- `phone_number_format` – The format to convert the phone number to. If no format is specified, the default is `E.164`, an internationally-recognized standard phone number format. Valid values include the following:
  - E164 (omit the period after E)
- `default_region` – A valid region code consisting of two or three uppercase letters that specifies the region for the phone number when no country code is present in the number itself. At most, one of `defaultRegion` or `defaultRegionColumn` can be provided.
- `default_region_column` – The name of a column of the advanced data type `Country`. The region code from the specified column is used to determine the country code for the phone number when no country code is present in the number itself. At most, one of `defaultRegion` or `defaultRegionColumn` can be provided.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

FormatCase class

The `FormatCase` transform changes each string in a column to the specified case type.

Example

```
from pyspark.context import SparkContext
from pyspark.sql import SparkSession
from awsgluedi.transforms import *

sc = SparkContext()
spark = SparkSession(sc)

datasource1 = spark.read.json("s3://${BUCKET}/json/zips/raw/data")

try:
    df_output = data_cleaning.FormatCase.apply(
        data_frame=datasource1,
        spark_context=sc,
        source_column="city",
        case_type="LOWER"
    )
except:
    print("Unexpected Error happened ")
    raise
```

Output

The `FormatCase` transformation will convert the values in the `city` column to lowercase based on the `case_type="LOWER"` parameter. The resulting `df_output` DataFrame will contain all columns from the original `datasource1` DataFrame, but with the `city` column values in lowercase.

Methods

- __call__
- apply
- name
- describeArgs
- describeReturn
- describeTransform
- describeErrors
- describe

__call__(spark_context, data_frame, source_column, case_type)

The `FormatCase` transform changes each string in a column to the specified case type.

- `source_column` – The name of an existing column.
- `case_type` – Supported case types are `CAPITAL`,`LOWER`, `UPPER`, `SENTENCE`.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

FillWithMode class

The `FillWithMode` transform formats a column according to the phone numberformat you specify. You can also specify tie-breaker logic, where some of the values are identical. For example, consider the following values: `1 2 2 3 3 4`

A modeType of `MINIMUM` causes `FillWithMode` to return 2 as the mode value. If modeType is `MAXIMUM`, the mode is 3. For `AVERAGE`, the mode is 2.5.

Example

```
from awsglue.context import *
from pyspark.sql import SparkSession
from awsgluedi.transforms import *

sc = SparkContext()
spark = SparkSession(sc)

input_df = spark.createDataFrame(
    [
        (105.111, 13.12),
        (1055.123, 13.12),
        (None, 13.12),
        (13.12, 13.12),
        (None, 13.12),
    ],
    ["source_column_1", "source_column_2"],
)

try:
    df_output = data_quality.FillWithMode.apply(
        data_frame=input_df,
        spark_context=sc,
        source_column="source_column_1",
        mode_type="MAXIMUM"
    )
    df_output.show()
except:
    print("Unexpected Error happened ")
    raise
```

Output

The output of the given code will be:

```
```
+---------------+---------------+
|source_column_1|source_column_2|
+---------------+---------------+
105.111	13.12
1055.123	13.12
1055.123	13.12
13.12	13.12
1055.123	13.12
+---------------+---------------+
```
```

The `FillWithMode` transformation from the `awsglue.data_quality` module is applied to the `input_df` DataFrame. It replaces the `null` values in the `source_column_1` column with the maximum value (`mode_type="MAXIMUM"`) from the non-null values in that column.

In this case, the maximum value in the `source_column_1` column is `1055.123`. Therefore, the `null` values in `source_column_1` are replaced by `1055.123` in the output DataFrame `df_output`.

Methods

- __call__
- apply
- name
- describeArgs
- describeReturn
- describeTransform
- describeErrors
- describe

__call__(spark_context, data_frame, source_column, mode_type)

The `FillWithMode` transform formats the case of strings in a column.

- `source_column` – The name of an existing column.
- `mode_type` – How to resolve tie values in the data. This value must be one of `MINIMUM`, `NONE`, `AVERAGE`, or `MAXIMUM`.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

FlagDuplicateRows class

The `FlagDuplicateRows` transform returns a new column with a specified value in each row that indicates whether that row is an exact match of an earlier row in the dataset. When matches are found, they are flagged as duplicates. The initial occurrence is not flagged, because it doesn't match an earlier row.

Example

```
from pyspark.context import SparkContext
from pyspark.sql import SparkSession
from awsgluedi.transforms import *

sc = SparkContext()
spark = SparkSession(sc)

input_df = spark.createDataFrame(
    [
        (105.111, 13.12),
        (13.12, 13.12),
        (None, 13.12),
        (13.12, 13.12),
        (None, 13.12),
    ],
    ["source_column_1", "source_column_2"],
)

try:
    df_output = data_quality.FlagDuplicateRows.apply(
        data_frame=input_df,
        spark_context=sc,
        target_column="flag_row",
        true_string="True",
        false_string="False",
        target_index=1
    )
except:
    print("Unexpected Error happened ")
    raise
```

Output

The output will be a PySpark DataFrame with an additional column `flag_row` that indicates whether a row is a duplicate or not, based on the `source_column_1` column. The resulting \`df_output\` DataFrame will contain the following rows:

```
```
+---------------+---------------+--------+
|source_column_1|source_column_2|flag_row|
```
```

```
+---------------+---------------+--------+
| 105.111| 13.12| False|
| 13.12| 13.12| True|
| null| 13.12| True|
| 13.12| 13.12| True|
| null| 13.12| True|
+---------------+---------------+--------+
```
```

The `flag_row` column indicates whether a row is a duplicate or not. The `true_string` is set to "True", and the `false_string` is set to "False". The `target_index` is set to 1, which means that the `flag_row` column will be inserted at the second position (index 1) in the output DataFrame.

#### Methods

- __call__
- apply
- name
- describeArgs
- describeReturn
- describeTransform
- describeErrors
- describe

### __call__(spark_context, data_frame, target_column, true_string=DEFAULT_TRUE_STRING, false_string=DEFAULT_FALSE_STRING, target_index=None)

The `FlagDuplicateRows` transform returns a new column with a specified value in each row that indicates whether that row is an exact match of an earlier row in the dataset. When matches are found, they are flagged as duplicates. The initial occurrence is not flagged, because it doesn't match an earlier row.

- `true_string` – Value to be inserted if the row matches an earlier row.
- `false_string` – Value to be inserted if the row is unique.
- `target_column` – Name of the new column that is inserted in the dataset.
```

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

RemoveDuplicates class

The `RemoveDuplicates` transform deletes an entire row, if a duplicate value is encountered in a selected source column.

Example

```
from pyspark.context import SparkContext
from pyspark.sql import SparkSession
from awsgluedi.transforms import *

sc = SparkContext()
spark = SparkSession(sc)

```

```
input_df = spark.createDataFrame(
    [
        (105.111, 13.12),
        (13.12, 13.12),
        (None, 13.12),
        (13.12, 13.12),
        (None, 13.12),
    ],
    ["source_column_1", "source_column_2"],
)

try:
    df_output = data_quality.RemoveDuplicates.apply(
        data_frame=input_df,
        spark_context=sc,
        source_column="source_column_1"
    )
except:
    print("Unexpected Error happened ")
    raise
```

Output

The output will be a PySpark DataFrame with duplicates removed based on the `source_column_1` column. The resulting `df_output` DataFrame will contain the following rows:

```
```

+---------------+---------------+
|source_column_1|source_column_2|
+---------------+---------------+
105.111	13.12
13.12	13.12
null	13.12
+---------------+---------------+
```
```

Note that the rows with `source_column_1` values of `13.12` and `null` appear only once in the output DataFrame, as the duplicates have been removed based on the `source_column_1` column.

Methods

- __call__

- apply
- name
- describeArgs
- describeReturn
- describeTransform
- describeErrors
- describe

__call__(spark_context, data_frame, source_column)

The `RemoveDuplicates` transform deletes an entire row, if a duplicate value is encountered in a selected source column.

- `source_column` – The name of an existing column.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

MonthName class

The `MonthName` transform creates a new column containing the name of the month, from a string that represents a date.

Example

```
from pyspark.context import SparkContext
from pyspark.sql import SparkSession
from awsgluedi.transforms import *

sc = SparkContext()
spark = SparkSession(sc)

spark.conf.set("spark.sql.legacy.timeParserPolicy", "LEGACY")

input_df = spark.createDataFrame(
    [
        ("20-2018-12",),
        ("2018-20-12",),
        ("20182012",),
        ("12202018",),
        ("20122018",),
        ("20-12-2018",),
        ("12/20/2018",),
        ("02/02/02",),
        ("02 02 2009",),
        ("02/02/2009",),
        ("August/02/2009",),
        ("02/june/2009",),
        ("02/2020/june",),
        ("2013-02-21 06:35:45.658505",),
        ("August 02 2009",),
        ("2013/02/21",),
        (None,),
    ],
    ["column_1"],
)

try:
    df_output = datetime_functions.MonthName.apply(
        data_frame=input_df,
        spark_context=sc,
        source_column="column_1",
        target_column="target_column"
    )
    df_output.show()
except:
    print("Unexpected Error happened ")
    raise
```

Output

The output will be:

```
```
+------------+------------+
| column_1|target_column|
+------------+------------+
20-2018-12	December
2018-20-12	null
20182012	null
12202018	null
20122018	null
20-12-2018	December
12/20/2018	December
02/02/02	February
02 02 2009	February
02/02/2009	February
August/02/2009	August
02/june/2009	null
02/2020/june	null
2013-02-21 06:35:45.658505	February
August 02 2009	August
2013/02/21	February
null	null
+------------+------------+
```
```

The `MonthName` transformation takes the `source_column` as `"column_1"` and the `target_column` as `"target_column"`. It attempts to extract the month name from the date/time strings in the `"column_1"` column and places it in the `"target_column"` column. If the date/

time string is in an unrecognized format or cannot be parsed, the `"target_column"` value is set to `null`.

The transformation successfully extracts the month name from various date/time formats, such as "20-12-2018", "12/20/2018", "02/02/2009", "2013-02-21 06:35:45.658505", and "August 02 2009".

Methods

- __call__
- apply
- name
- describeArgs
- describeReturn
- describeTransform
- describeErrors
- describe

__call__(spark_context, data_frame, target_column, source_column=None, value=None)

The `MonthName` transform creates a new column containing the name of the month, from a string that represents a date.

- `source_column` – The name of an existing column.
- `value` – A character string to evaluate..
- `target_column` – A name for the newly created column.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

IsEven class

The `IsEven` transform returns a Boolean value in a new column that indicates whether the source column or value is even. If the source column or value is a decimal, the result is false.

Example

```
from pyspark.context import SparkContext
from pyspark.sql import SparkSession
from awsgluedi.transforms import *

sc = SparkContext()
spark = SparkSession(sc)

input_df = spark.createDataFrame(
    [(5,), (0,), (-1,), (2,), (None,)],
    ["source_column"],
)

try:
    df_output = math_functions.IsEven.apply(
        data_frame=input_df,
        spark_context=sc,
        source_column="source_column",
        target_column="target_column",
        value=None,
        true_string="Even",
        false_string="Not even",
    )
    df_output.show()
except:
    print("Unexpected Error happened ")
    raise
```

Output

The output will be:

```
```
+------------+------------+
|source_column|target_column|
+------------+------------+
5	Not even
0	Even
-1	Not even
2	Even
null	null
+------------+------------+
```
```

The `IsEven` transformation takes the `source_column` as "source_column" and the `target_column` as "target_column". It checks if the value in the `"source_column"` is even or not. If the value is even, it sets the `"target_column"` value to the `true_string` "Even". If the value is odd, it sets the `"target_column"` value to the `false_string` "Not even". If the `"source_column"` value is `null`, the `"target_column"` value is set to `null`.

The transformation correctly identifies the even numbers (0 and 2) and sets the `"target_column"` value to "Even". For odd numbers (5 and -1), it sets the `"target_column"` value to "Not even". For the `null` value in `"source_column"`, the `"target_column"` value is set to `null`.

Methods

- __call__
- apply
- name
- describeArgs

- describeReturn
- describeTransform
- describeErrors
- describe

__call__(spark_context, data_frame, target_column, source_column=None, true_string=DEFAULT_TRUE_STRING, false_string=DEFAULT_FALSE_STRING, value=None)

The `IsEven` transform returns a Boolean value in a new column that indicates whether the source column or value is even. If the source column or value is a decimal, the result is false.

- `source_column` – The name of an existing column.
- `target_column` – The name of the new column to be created.
- `true_string` – A string that indicates whether the value is even.
- `false_string` – A string that indicates whether the value is not even.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

CryptographicHash class

The `CryptographicHash` transform applies an algorithm to hash values in the column.

Example

```
from pyspark.context import SparkContext
from pyspark.sql import SparkSession
from awsgluedi.transforms import *

secret = "${SECRET}"
sc = SparkContext()
spark = SparkSession(sc)

input_df = spark.createDataFrame(
    [
        (1, "1234560000"),
        (2, "1234560001"),
        (3, "1234560002"),
        (4, "1234560003"),
        (5, "1234560004"),
        (6, "1234560005"),
        (7, "1234560006"),
        (8, "1234560007"),
        (9, "1234560008"),
        (10, "1234560009"),
    ],
    ["id", "phone"],
)

try:
    df_output = pii.CryptographicHash.apply(
        data_frame=input_df,
        spark_context=sc,
        source_columns=["id", "phone"],
        secret_id=secret,
        algorithm="HMAC_SHA256",
        output_format="BASE64",
    )
    df_output.show()
```

```
except:
    print("Unexpected Error happened ")
    raise
```

Output

The output will be:

~~~
```

+---+------------+-------------------+-------------------+
| id| phone | id_hashed | phone_hashed |
+---+------------+-------------------+-------------------+
| 1| 1234560000 | QUI1zXTJiXmfIb... | juDBAmiRnnO3g... |
| 2| 1234560001 | ZAUWiZ3dVTzCo... | vC8lgUqBVDMNQ... |
| 3| 1234560002 | ZP4VvZWkqYifu... | Kl3QAkgswYpzB... |
| 4| 1234560003 | 3u8vO3wQ8EQfj... | CPBzK1P8PZZkV... |
| 5| 1234560004 | eWkQJk4zAOIzx... | aLf7+mHcXqbLs... |
| 6| 1234560005 | xtI9fZCJZCvsa... | dy2DFgdYWmr0p... |
| 7| 1234560006 | iW9hew7jnHuOf... | wwfGMCOEv6oOv... |
| 8| 1234560007 | H9V1pqvgkFhfS... | g9WKhagIXy9ht... |
| 9| 1234560008 | xDhEuHaxAUbU5... | b3uQLKPY+Q5vU... |
| 10| 1234560009 | GRN6nFXkxk349... | VJdsKt8VbxBbt... |
+---+------------+-------------------+-------------------+
```
~~~

The transformation computes the cryptographic hashes of the values in the `id` and `phone` columns using the specified algorithm and secret key, and encodes the hashes in Base64 format. The resulting `df_output` DataFrame contains all columns from the original `input_df` DataFrame, plus the additional `id_hashed` and `phone_hashed` columns with the computed hashes.

Methods

- __call__
- apply
- name
- describeArgs
- describeReturn
- describeTransform

- describeErrors
- describe

__call__(spark_context, data_frame, source_columns, secret_id, algorithm=None, secret_version=None, create_secret_if_missing=False, output_format=None, entity_type_filter=None)

The `CryptographicHash` transform applies an algorithm to hash values in the column.

- `source_columns` – An array of existing columns.
- `secret_id` – The ARN of the Secrets Manager secret key. The key used in the hash-based message authentication code (HMAC) prefix algorithm to hash the source columns.
- `secret_version` – Optional. Defaults to the latest secret version.
- `entity_type_filter` – Optional array of entity types. Can be used to encrypt only detected PII in free-text column.
- `create_secret_if_missing` – Optional boolean. If true will attempt to create the secret on behalf of the caller.
- `algorithm` – The algorithm used to hash your data. Valid enum values: MD5, SHA1, SHA256, SHA512, HMAC_MD5, HMAC_SHA1, HMAC_SHA256, HMAC_SHA512.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

Decrypt class

The `Decrypt` transform decrypts inside of AWS Glue. Your data can also be decrypted outside of AWS Glue with the AWS Encryption SDK. If the provided KMS key ARN does not match what was used to encrypt the column, the decrypt operation fails.

Example

```
from pyspark.context import SparkContext
from pyspark.sql import SparkSession
from awsgluedi.transforms import *

kms = "${KMS}"
sc = SparkContext()
spark = SparkSession(sc)

input_df = spark.createDataFrame(
    [
        (1, "1234560000"),
        (2, "1234560001"),
        (3, "1234560002"),
        (4, "1234560003"),
        (5, "1234560004"),
        (6, "1234560005"),
        (7, "1234560006"),
        (8, "1234560007"),
        (9, "1234560008"),
        (10, "1234560009"),
    ],
    ["id", "phone"],
)

try:
    df_encrypt = pii.Encrypt.apply(
        data_frame=input_df,
        spark_context=sc,
        source_columns=["phone"],
        kms_key_arn=kms
    )
    df_decrypt = pii.Decrypt.apply(
        data_frame=df_encrypt,
        spark_context=sc,
        source_columns=["phone"],
        kms_key_arn=kms
    )
    df_decrypt.show()
except:
    print("Unexpected Error happened ")
    raise
```

Output

The output will be a PySpark DataFrame with the original `id` column and the decrypted `phone` column:

~~~
```
+---+------------+
| id| phone|
+---+------------+
| 1| 1234560000|
| 2| 1234560001|
| 3| 1234560002|
| 4| 1234560003|
| 5| 1234560004|
| 6| 1234560005|
| 7| 1234560006|
| 8| 1234560007|
| 9| 1234560008|
| 10| 1234560009|
+---+------------+
```
~~~

The `Encrypt` transform takes the `source_columns` as `["phone"]` and the `kms_key_arn` as the value of the `${KMS}` environment variable. The transformation encrypts the values in the `phone` column using the specified KMS key. The encrypted DataFrame `df_encrypt` is then passed to the `Decrypt` transform from the `awsglue.pii` module. It takes the `source_columns` as `["phone"]` and the `kms_key_arn` as the value of the `${KMS}` environment variable. The transformation decrypts the encrypted values in the `phone` column using the same KMS key. The resulting `df_decrypt` DataFrame contains the original `id` column and the decrypted `phone` column.

Methods

- __call__
- apply
- name
- describeArgs
- describeReturn
- describeTransform
- describeErrors
- describe

__call__(spark_context, data_frame, source_columns, kms_key_arn)

The `Decrypt` transform decrypts inside of AWS Glue. Your data can also be decrypted outside of AWS Glue with the AWS Encryption SDK. If the provided KMS key ARN does not match what was used to encrypt the column, the decrypt operation fails.

- `source_columns` – An array of existing columns.
- `kms_key_arn` – The key ARN of the AWS Key Management Service key to use to decrypt the source columns.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

Encrypt class

The `Encrypt` transform encrypts source columns using the AWS Key Management Service key. The `Encrypt` transform can encrypt up to 128 MiB per cell. It will attempt to preserve the format on decryption. To preserve the data type, the data type metadata must serialize to less than 1KB. Otherwise, you must set the `preserve_data_type` parameter to false. The data type metadata will be stored in plaintext in the encryption context.

Example

```
from pyspark.context import SparkContext
from pyspark.sql import SparkSession
from awsgluedi.transforms import *

kms = "${KMS}"
sc = SparkContext()
spark = SparkSession(sc)

input_df = spark.createDataFrame(
    [
        (1, "1234560000"),
        (2, "1234560001"),
        (3, "1234560002"),
        (4, "1234560003"),
        (5, "1234560004"),
        (6, "1234560005"),
        (7, "1234560006"),
        (8, "1234560007"),
        (9, "1234560008"),
        (10, "1234560009"),
    ],
    ["id", "phone"],
)

try:
    df_encrypt = pii.Encrypt.apply(
        data_frame=input_df,
        spark_context=sc,
        source_columns=["phone"],
        kms_key_arn=kms
    )
except:
    print("Unexpected Error happened ")
    raise
```

Output

The output will be a PySpark DataFrame with the original `id` column and an additional column containing the encrypted values of the `phone` column.

~~~
```
+---+------------+-------------------------+
| id| phone | phone_encrypted |
+---+------------+-------------------------+
| 1| 1234560000| EncryptedData1234...abc |
| 2| 1234560001| EncryptedData5678...def |
| 3| 1234560002| EncryptedData9012...ghi |
| 4| 1234560003| EncryptedData3456...jkl |
| 5| 1234560004| EncryptedData7890...mno |
| 6| 1234560005| EncryptedData1234...pqr |
| 7| 1234560006| EncryptedData5678...stu |
| 8| 1234560007| EncryptedData9012...vwx |
| 9| 1234560008| EncryptedData3456...yz0 |
| 10| 1234560009| EncryptedData7890...123 |
+---+------------+-------------------------+
```
~~~

The `Encrypt` transform takes the `source_columns` as `["phone"]` and the `kms_key_arn` as the value of the `${KMS}` environment variable. The transformation encrypts the values in the `phone` column using the specified KMS key. The resulting `df_encrypt` DataFrame contains the original `id` column, the original `phone` column, and an additional column named `phone_encrypted` containing the encrypted values of the `phone` column.

Methods

- __call__
- apply
- name
- describeArgs
- describeReturn
- describeTransform
- describeErrors
- describe

__call__(spark_context, data_frame, source_columns, kms_key_arn, entity_type_filter=None, preserve_data_type=None)

The `Encrypt` transform encrypts source columns using the AWS Key Management Service key.

- `source_columns` – An array of existing columns.
- `kms_key_arn` – The key ARN of the AWS Key Management Service key to use to Encrypt the source columns.
- `entity_type_filter` – Optional array of entity types. Can be used to encrypt only detected PII in free-text column.
- `preserve_data_type` – Optional boolean. Defaults to true. If false, the data type will not be stored.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

IntToIp class

The `IntToIp` transform converts the integer value of source column or other value to the corresponding IPv4 value in then target column, and returns the result in a new column.

Example

```
from pyspark.context import SparkContext
from pyspark.sql import SparkSession
from awsgluedi.transforms import *

sc = SparkContext()
spark = SparkSession(sc)

input_df = spark.createDataFrame(
    [
        (3221225473,),
        (0,),
        (1,),
        (100,),
        (168430090,),
        (4294967295,),
        (4294967294,),
        (4294967296,),
        (-1,),
        (None,),
    ],
    ["source_column_int"],
)

try:
    df_output = web_functions.IntToIp.apply(
        data_frame=input_df,
        spark_context=sc,
        source_column="source_column_int",
        target_column="target_column",
        value=None
    )
    df_output.show()
except:
    print("Unexpected Error happened ")
    raise
```

Output

The output will be:

```
```
+---------------+---------------+
|source_column_int|target_column|
+---------------+---------------+
3221225473	192.0.0.1
0	0.0.0.0
1	0.0.0.1
100	0.0.0.100
168430090	10.0.0.10
4294967295	255.255.255.255
4294967294	255.255.255.254
4294967296	null
-1	null
null	null
+---------------+---------------+
```
```

The `IntToIp.apply` transformation takes the `source_column` as `"source_column_int"` and the `target_column` as `"target_column"` and converts the integer values in the `source_column_int` column to their corresponding IPv4 address representation and stores the result in the `target_column` column.

For valid integer values within the range of IPv4 addresses (0 to 4294967295), the transformation successfully converts them to their IPv4 address representation (e.g., 192.0.0.1, 0.0.0.0, 10.0.0.10, 255.255.255.255).

For integer values outside the valid range (e.g., 4294967296, -1), the `target_column` value is set to `null`. For `null` values in the `source_column_int` column, the `target_column` value is also set to `null`.

Methods

- __call__
- apply
- name
- describeArgs
- describeReturn
- describeTransform
- describeErrors
- describe

__call__(spark_context, data_frame, target_column, source_column=None, value=None)

The `IntToIp` transform converts the integer value of source column or other value to the corresponding IPv4 value in then target column, and returns the result in a new column.

- `sourceColumn` – The name of an existing column.
- `value` – A character string to evaluate.
- `targetColumn` – The name of the new column to be created.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

IpToInt class

The `IpToInt` transform converts the Internet Protocol version 4 (IPv4) value of the source column or other value to the corresponding integer value in the target column, and returns the result in a new column.

Example

For AWS Glue 4.0 and above, create or update job arguments with `key: --enable-glue-di-transforms, value: true`

```
from pyspark.context import SparkContext
from awsgluedi.transforms import *

sc = SparkContext()

input_df = spark.createDataFrame(
    [
        ("192.0.0.1",),
        ("10.10.10.10",),
        ("1.2.3.4",),
        ("1.2.3.6",),
        ("http://12.13.14.15",),
        ("https://16.17.18.19",),
        ("1.2.3.4",),
        (None,),
        ("abc",),
        ("abc.abc.abc.abc",),
        ("321.123.123.123",),
        ("244.4.4.4",),
        ("255.255.255.255",),
    ],
    ["source_column_ip"],
)

    df_output = web_functions.IpToInt.apply(
        data_frame=input_df,
        spark_context=sc,
        source_column="source_column_ip",
        target_column="target_column",
        value=None
    )
    df_output.show()
```

Output

The output will be:

```
```
+----------------+---------------+
|source_column_ip| target_column|
+----------------+---------------+
192.0.0.1	3221225473
10.10.10.10	168427722
1.2.3.4	16909060
1.2.3.6	16909062
http://12.13.14.15	null
https://16.17.18.19	null
1.2.3.4	16909060
null	null
abc	null
abc.abc.abc.abc	null
321.123.123.123	null
244.4.4.4	4102444804
255.255.255.255	4294967295
+----------------+---------------+
```
```

The `IpToInt` transformation takes the `source_column` as `"source_column_ip"` and the `target_column` as `"target_column"` and converts the valid IPv4 address strings in the `source_column_ip` column to their corresponding 32-bit integer representation and stores the result in the `target_column` column.

For valid IPv4 address strings (e.g., "192.0.0.1", "10.10.10.10", "1.2.3.4"), the transformation successfully converts them to their integer representation (e.g., 3221225473, 168427722, 16909060). For strings that are not valid IPv4 addresses (e.g., URLs, non-IP strings like "abc", invalid IP formats like "abc.abc.abc.abc"), the `target_column` value is set to `null`. For `null` values in the `source_column_ip` column, the `target_column` value is also set to `null`.

Methods

- __call__
- apply
- name
- describeArgs
- describeReturn
- describeTransform
- describeErrors
- describe

__call__(spark_context, data_frame, target_column, source_column=None, value=None)

The `IpToInt` transform converts the Internet Protocol version 4 (IPv4) value of the source column or other value to the corresponding integer value in the target column, and returns the result in a new column.

- `sourceColumn` – The name of an existing column.
- `value` – A character string to evaluate.
- `targetColumn` – The name of the new column to be created.

apply(cls, *args, **kwargs)

Inherited from `GlueTransform` apply.

name(cls)

Inherited from `GlueTransform` name.

describeArgs(cls)

Inherited from `GlueTransform` describeArgs.

describeReturn(cls)

Inherited from `GlueTransform` describeReturn.

describeTransform(cls)

Inherited from `GlueTransform` describeTransform.

describeErrors(cls)

Inherited from `GlueTransform` describeErrors.

describe(cls)

Inherited from `GlueTransform` describe.

Data integration transforms

For AWS Glue 4.0 and above, create or update job arguments with `key: --enable-glue-di-transforms, value: true`.

Example job script:

```
from pyspark.context import SparkContext

from awsgluedi.transforms import *
sc = SparkContext()

input_df = spark.createDataFrame(
    [(5,), (0,), (-1,), (2,), (None,)],
    ["source_column"],
)

try:
    df_output = math_functions.IsEven.apply(
        data_frame=input_df,
        spark_context=sc,
        source_column="source_column",
        target_column="target_column",
        value=None,
        true_string="Even",
        false_string="Not even",
    )
    df_output.show()
except:
    print("Unexpected Error happened ")
    raise
```

Example Sessions using Notebooks

```
%idle_timeout 2880
%glue_version 4.0
%worker_type G.1X
%number_of_workers 5
%region eu-west-1
```

```
%%configure
{
    "--enable-glue-di-transforms": "true"
}
```

```
from pyspark.context import SparkContext
from awsgluedi.transforms import *

sc = SparkContext()

input_df = spark.createDataFrame(
    [(5,), (0,), (-1,), (2,), (None,)],
    ["source_column"],
)

try:
    df_output = math_functions.IsEven.apply(
        data_frame=input_df,
        spark_context=sc,
        source_column="source_column",
        target_column="target_column",
        value=None,
        true_string="Even",
        false_string="Not even",
    )
    df_output.show()
except:
    print("Unexpected Error happened ")
    raise
```

Example Sessions using AWS CLI

```
aws glue create-session --default-arguments "--enable-glue-di-transforms=true"
```

DI transforms:

- FlagDuplicatesInColumn class
- FormatPhoneNumber class
- FormatCase class
- FillWithMode class
- FlagDuplicateRows class
- RemoveDuplicates class
- MonthName class
- IsEven class
- CryptographicHash class
- Decrypt class
- Encrypt class
- IntToIp class
- IpToInt class

Maven: Bundle the plugin with your Spark applications

You can bundle the transforms dependency with your Spark applications and Spark distributions (version 3.3) by adding the plugin dependency in your Maven `pom.xml` while developing your Spark applications locally.

```
<repositories>
   ...
    <repository>
        <id>aws-glue-etl-artifacts</id>
        <url>https://aws-glue-etl-artifacts.s3.amazonaws.com/release/ </url>
    </repository>
</repositories>
...
<dependency>
    <groupId>com.amazonaws</groupId>
    <artifactId>AWSGlueTransforms</artifactId>
    <version>4.0.0</version>
</dependency>
```

You can alternatively download the binaries from AWS Glue Maven artifacts directly and include them in your Spark application as follows.

```
#!/bin/bash
sudo wget -v https://aws-glue-etl-artifacts.s3.amazonaws.com/release/com/amazonaws/
AWSGlueTransforms/4.0.0/AWSGlueTransforms-4.0.0.jar -P /usr/lib/spark/jars/
```

Programming AWS Glue ETL scripts in Scala

You can find Scala code examples and utilities for AWS Glue in the AWS Glue samples repository on the GitHub website.

AWS Glue supports an extension of the PySpark Scala dialect for scripting extract, transform, and load (ETL) jobs. The following sections describe how to use the AWS Glue Scala library and the AWS Glue API in ETL scripts, and provide reference documentation for the library.

Contents

- Using Scala to program AWS Glue ETL scripts
  - Testing a Scala ETL program in a Jupyter notebook on a development endpoint
  - Testing a Scala ETL program in a Scala REPL
- Scala script example - streaming ETL
- APIs in the AWS Glue Scala library
  - com.amazonaws.services.glue
  - com.amazonaws.services.glue.ml
  - com.amazonaws.services.glue.dq
  - com.amazonaws.services.glue.types
  - com.amazonaws.services.glue.util
  - AWS Glue Scala ChoiceOption APIs

    - ChoiceOption trait
    - ChoiceOption object
        - Def apply
    - Case class ChoiceOptionWithResolver
    - Case class MatchCatalogSchemaChoiceOption
- Abstract DataSink class
    - Def writeDynamicFrame
    - Def pyWriteDynamicFrame
    - Def writeDataFrame
    - Def pyWriteDataFrame
    - Def setCatalogInfo
    - Def supportsFormat
    - Def setFormat
    - Def withFormat
    - Def setAccumulableSize
    - Def getOutputErrorRecordsAccumulable
    - Def errorsAsDynamicFrame
    - DataSink object
        - Def recordMetrics
- AWS Glue Scala DataSource trait
- AWS Glue Scala DynamicFrame APIs
    - AWS Glue Scala DynamicFrame class
        - Val errorsCount
        - Def applyMapping
        - Def assertErrorThreshold
        - Def count
        - Def dropField
        - Def dropFields
        - Def dropNulls
        - Def errorsAsDynamicFrame

- Def filter
- Def getName
- Def getNumPartitions
- Def getSchemaIfComputed
- Def isSchemaComputed
- Def javaToPython
- Def join
- Def map
- Def mergeDynamicFrames
- Def printSchema
- Def recomputeSchema
- Def relationalize
- Def renameField
- Def repartition
- Def resolveChoice
- Def schema
- Def selectField
- Def selectFields
- Def show
- Def simplifyDDBJson
- Def spigot
- Def splitFields
- Def splitRows
- Def stageErrorsCount
- Def toDF
- Def unbox
- Def unnest
- Def unnestDDBJson
- Def withFrameSchema
- Def withName

    - Def withTransformationContext
  - The DynamicFrame object
    - Def apply
    - Def emptyDynamicFrame
    - Def fromPythonRDD
    - Def ignoreErrors
    - Def inlineErrors
    - Def newFrameWithErrors
- AWS Glue Scala DynamicRecord class
  - Def addField
  - Def dropField
  - Def setError
  - Def isError
  - Def getError
  - Def clearError
  - Def write
  - Def readFields
  - Def clone
  - Def schema
  - Def getRoot
  - Def toJson
  - Def getFieldNode
  - Def getField
  - Def hashCode
  - Def equals
  - DynamicRecord object
    - Def apply
  - RecordTraverser trait
- AWS Glue Scala GlueContext APIs
  - def addIngestionTimeColumns

  - def createDataFrameFromOptions
  - forEachBatch
  - def getCatalogSink
  - def getCatalogSource
  - def getJDBCSink
  - def getSink
  - def getSinkWithFormat
  - def getSource
  - def getSourceWithFormat
  - def getSparkSession
  - def startTransaction
  - def commitTransaction
  - def cancelTransaction
  - def this
  - def this
  - def this
- MappingSpec
  - MappingSpec case class
  - MappingSpec object
  - Val orderingByTarget
  - Def apply
  - Def apply
  - Def apply
- AWS Glue Scala ResolveSpec APIs
  - ResolveSpec object
    - Def
    - Def
  - ResolveSpec case class
    - ResolveSpec def methods
- AWS Glue Scala ArrayNode APIs

  - ArrayNode case class
    - ArrayNode def methods
- AWS Glue Scala BinaryNode APIs
  - BinaryNode case class
    - BinaryNode val fields
    - BinaryNode def methods
- AWS Glue Scala BooleanNode APIs
  - BooleanNode case class
    - BooleanNode val fields
    - BooleanNode def methods
- AWS Glue Scala ByteNode APIs
  - ByteNode case class
    - ByteNode val fields
    - ByteNode def methods
- AWS Glue Scala DateNode APIs
  - DateNode case class
    - DateNode val fields
    - DateNode def methods
- AWS Glue Scala DecimalNode APIs
  - DecimalNode case class
    - DecimalNode val fields
    - DecimalNode def methods
- AWS Glue Scala DoubleNode APIs
  - DoubleNode case class
    - DoubleNode val fields
    - DoubleNode def methods
- AWS Glue Scala DynamicNode APIs
  - DynamicNode class
    - DynamicNode def methods
  - DynamicNode object

        - DynamicNode def methods
- EvaluateDataQuality class
    - Def apply
    - Example
- AWS Glue Scala FloatNode APIs
    - FloatNode case class
        - FloatNode val fields
        - FloatNode def methods
- FillMissingValues class
    - Def apply
- FindMatches class
    - Def apply
- FindIncrementalMatches class
    - Def apply
- AWS Glue Scala IntegerNode APIs
    - IntegerNode case class
        - IntegerNode val fields
        - IntegerNode def methods
- AWS Glue Scala LongNode APIs
    - LongNode case class
        - LongNode val fields
        - LongNode def methods
- AWS Glue Scala MapLikeNode APIs
    - MapLikeNode class
        - MapLikeNode def methods
- AWS Glue Scala MapNode APIs
    - MapNode case class
        - MapNode def methods
- AWS Glue Scala NullNode APIs

    - NullNode class

    - NullNode case object
- AWS Glue Scala ObjectNode APIs
    - ObjectNode object
        - ObjectNode def methods
    - ObjectNode case class
        - ObjectNode def methods
- AWS Glue Scala ScalarNode APIs
    - ScalarNode class
        - ScalarNode def methods
    - ScalarNode object
        - ScalarNode def methods
- AWS Glue Scala ShortNode APIs
    - ShortNode case class
        - ShortNode val fields
        - ShortNode def methods
- AWS Glue Scala StringNode APIs
    - StringNode case class
        - StringNode val fields
        - StringNode def methods
- AWS Glue Scala TimestampNode APIs
    - TimestampNode case class
        - TimestampNode val fields
        - TimestampNode def methods
- AWS Glue Scala GlueArgParser APIs
    - GlueArgParser object
        - GlueArgParser def methods
- AWS Glue Scala job APIs
    - Job object
        - Job def methods

Using Scala to program AWS Glue ETL scripts

You can automatically generate a Scala extract, transform, and load (ETL) program using the AWS Glue console, and modify it as needed before assigning it to a job. Or, you can write your own program from scratch. For more information, see Configuring job properties for Spark jobs in AWS Glue. AWS Glue then compiles your Scala program on the server before running the associated job.

To ensure that your program compiles without errors and runs as expected, it's important that you load it on a development endpoint in a REPL (Read-Eval-Print Loop) or a Jupyter Notebook and test it there before running it in a job. Because the compile process occurs on the server, you will not have good visibility into any problems that happen there.

Testing a Scala ETL program in a Jupyter notebook on a development endpoint

To test a Scala program on an AWS Glue development endpoint, set up the development endpoint as described in Adding a development endpoint.

Next, connect it to a Jupyter Notebook that is either running locally on your machine or remotely on an Amazon EC2 notebook server. To install a local version of a Jupyter Notebook, follow the instructions in Tutorial: Jupyter notebook in JupyterLab.

The only difference between running Scala code and running PySpark code on your Notebook is that you should start each paragraph on the Notebook with the the following:

```
%spark
```

This prevents the Notebook server from defaulting to the PySpark flavor of the Spark interpreter.

Testing a Scala ETL program in a Scala REPL

You can test a Scala program on a development endpoint using the AWS GlueScala REPL. Follow the instructions in Tutorial: Use a SageMaker AI notebook, except at the end of the SSH-to-REPL command, replace `-t gluepyspark` with `-t glue-spark-shell`. This invokes the AWS Glue Scala REPL.

To close the REPL when you are finished, type `sys.exit`.

Scala script example - streaming ETL

Example

The following example script connects to Amazon Kinesis Data Streams, uses a schema from the Data Catalog to parse a data stream, joins the stream to a static dataset on Amazon S3, and outputs the joined results to Amazon S3 in parquet format.

```
// This script connects to an Amazon Kinesis stream, uses a schema from the data
 catalog to parse the stream,
// joins the stream to a static dataset on Amazon S3, and outputs the joined results to
 Amazon S3 in parquet format.
import com.amazonaws.services.glue.GlueContext
import com.amazonaws.services.glue.util.GlueArgParser
import com.amazonaws.services.glue.util.Job
import java.util.Calendar
import org.apache.spark.SparkContext
import org.apache.spark.sql.Dataset
import org.apache.spark.sql.Row
import org.apache.spark.sql.SaveMode
import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.functions.from_json
import org.apache.spark.sql.streaming.Trigger
import scala.collection.JavaConverters._

object streamJoiner {
  def main(sysArgs: Array[String]) {
    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)
    val sparkSession: SparkSession = glueContext.getSparkSession
    import sparkSession.implicits._
    // @params: [JOB_NAME]
    val args = GlueArgParser.getResolvedOptions(sysArgs, Seq("JOB_NAME").toArray)
    Job.init(args("JOB_NAME"), glueContext, args.asJava)

    val staticData = sparkSession.read          // read() returns type DataFrameReader
      .format("csv")
      .option("header", "true")
      .load("s3://amzn-s3-demo-bucket/inputs/productsStatic.csv")  // load() returns a
 DataFrame

    val datasource0 = sparkSession.readStream   // readstream() returns type
 DataStreamReader
```

```
      .format("kinesis")
      .option("streamName", "stream-join-demo")
      .option("endpointUrl", "https://kinesis.us-east-1.amazonaws.com")
      .option("startingPosition", "TRIM_HORIZON")
      .load                                     // load() returns a DataFrame

    val selectfields1 = datasource0.select(from_json($"data".cast("string"),
 glueContext.getCatalogSchemaAsSparkSchema("stream-demos", "stream-join-demo2")) as
 "data").select("data.*")

    val datasink2 = selectfields1.writeStream.foreachBatch { (dataFrame: Dataset[Row],
 batchId: Long) => {   //foreachBatch() returns type DataStreamWriter
      val joined = dataFrame.join(staticData, "product_id")
      val year: Int = Calendar.getInstance().get(Calendar.YEAR)
      val month :Int = Calendar.getInstance().get(Calendar.MONTH) + 1
      val day: Int = Calendar.getInstance().get(Calendar.DATE)
      val hour: Int = Calendar.getInstance().get(Calendar.HOUR_OF_DAY)

      if (dataFrame.count() > 0) {
        joined.write                           // joined.write returns type
 DataFrameWriter
          .mode(SaveMode.Append)
          .format("parquet")
          .option("quote", " ")
          .save("s3://amzn-s3-demo-bucket/output/" + "/year=" + "%04d".format(year)
 + "/month=" + "%02d".format(month) + "/day=" + "%02d".format(day) + "/hour=" +
 "%02d".format(hour) + "/")
      }
    }
    }  // end foreachBatch()
      .trigger(Trigger.ProcessingTime("100 seconds"))
      .option("checkpointLocation", "s3://amzn-s3-demo-bucket/checkpoint/")
      .start().awaitTermination()              // start() returns type StreamingQuery
    Job.commit()
  }
}
```

APIs in the AWS Glue Scala library

AWS Glue supports an extension of the PySpark Scala dialect for scripting extract, transform, and load (ETL) jobs. The following sections describe the APIs in the AWS Glue Scala library.

com.amazonaws.services.glue

The **com.amazonaws.services.glue** package in the AWS Glue Scala library contains the following APIs:

- ChoiceOption
- DataSink
- DataSource trait
- DynamicFrame
- DynamicRecord
- GlueContext
- MappingSpec
- ResolveSpec

com.amazonaws.services.glue.ml

The **com.amazonaws.services.glue.ml** package in the AWS Glue Scala library contains the following APIs:

- FillMissingValues
- FindIncrementalMatches
- FindMatches

com.amazonaws.services.glue.dq

The **com.amazonaws.services.glue.dq** package in the AWS Glue Scala library contains the following APIs:

- EvaluateDataQuality

com.amazonaws.services.glue.types

The **com.amazonaws.services.glue.types** package in the AWS Glue Scala library contains the following APIs:

- ArrayNode
- BinaryNode

- BooleanNode
- ByteNode
- DateNode
- DecimalNode
- DoubleNode
- DynamicNode
- FloatNode
- IntegerNode
- LongNode
- MapLikeNode
- MapNode
- NullNode
- ObjectNode
- ScalarNode
- ShortNode
- StringNode
- TimestampNode

com.amazonaws.services.glue.util

The **com.amazonaws.services.glue.util** package in the AWS Glue Scala library contains the following APIs:

- GlueArgParser
- Job

AWS Glue Scala ChoiceOption APIs

Topics

- ChoiceOption trait
- ChoiceOption object
- Case class ChoiceOptionWithResolver
- Case class MatchCatalogSchemaChoiceOption

Package: com.amazonaws.services.glue

ChoiceOption trait

```
trait ChoiceOption extends Serializable
```

ChoiceOption object

ChoiceOption

```
object ChoiceOption
```

A general strategy to resolve choice applicable to all `ChoiceType` nodes in a `DynamicFrame`.

- `val CAST`
- `val MAKE_COLS`
- `val MAKE_STRUCT`
- `val MATCH_CATALOG`
- `val PROJECT`

Def apply

```
def apply(choice: String): ChoiceOption
```

Case class ChoiceOptionWithResolver

```
case class ChoiceOptionWithResolver(name: String, choiceResolver: ChoiceResolver)
 extends ChoiceOption {}
```

Case class MatchCatalogSchemaChoiceOption

```
case class MatchCatalogSchemaChoiceOption() extends ChoiceOption {}
```

Abstract DataSink class

Topics

- Def writeDynamicFrame

- Def pyWriteDynamicFrame
- Def writeDataFrame
- Def pyWriteDataFrame
- Def setCatalogInfo
- Def supportsFormat
- Def setFormat
- Def withFormat
- Def setAccumulableSize
- Def getOutputErrorRecordsAccumulable
- Def errorsAsDynamicFrame
- DataSink object

Package: com.amazonaws.services.glue

```
abstract class DataSink
```

The writer analog to a `DataSource`. `DataSink` encapsulates a destination and a format that a `DynamicFrame` can be written to.

Def writeDynamicFrame

```
def writeDynamicFrame( frame : DynamicFrame,
                       callSite : CallSite = CallSite("Not provided", "")
                     ) : DynamicFrame
```

Def pyWriteDynamicFrame

```
def pyWriteDynamicFrame( frame : DynamicFrame,
                         site : String = "Not provided",
                         info : String = "" )
```

Def writeDataFrame

```
def writeDataFrame(frame: DataFrame,
                   glueContext: GlueContext,
                    callSite: CallSite = CallSite("Not provided", "")
                    ): DataFrame
```

Def pyWriteDataFrame

```
def pyWriteDataFrame(frame: DataFrame,
                     glueContext: GlueContext,
                     site: String = "Not provided",
                     info: String = ""
                     ): DataFrame
```

Def setCatalogInfo

```
def setCatalogInfo(catalogDatabase: String,
                   catalogTableName : String,
                   catalogId : String = "")
```

Def supportsFormat

```
def supportsFormat( format : String ) : Boolean
```

Def setFormat

```
def setFormat( format : String,
               options : JsonOptions
             ) : Unit
```

Def withFormat

```
def withFormat( format : String,
                options : JsonOptions = JsonOptions.empty
              ) : DataSink
```

Def setAccumulableSize

```
def setAccumulableSize( size : Int ) : Unit
```

Def getOutputErrorRecordsAccumulable

```
def getOutputErrorRecordsAccumulable : Accumulable[List[OutputError], OutputError]
```

Def errorsAsDynamicFrame

```
def errorsAsDynamicFrame : DynamicFrame
```

DataSink object

```
object DataSink
```

Def recordMetrics

```
def recordMetrics( frame : DynamicFrame,
                   ctxt : String
                 ) : DynamicFrame
```

AWS Glue Scala DataSource trait

Package: com.amazonaws.services.glue

A high-level interface for producing a `DynamicFrame`.

```
trait DataSource {

  def getDynamicFrame : DynamicFrame

  def getDynamicFrame( minPartitions : Int,
                       targetPartitions : Int
                     ) : DynamicFrame
  def getDataFrame : DataFrame

  /** @param num: the number of records for sampling.
    * @param options: optional parameters to control sampling behavior. Current
 available parameter for Amazon S3 sources in options:
    *  1. maxSamplePartitions: the maximum number of partitions the sampling will
 read.
    *  2. maxSampleFilesPerPartition: the maximum number of files the sampling will
 read in one partition.
```

```
    */
  def getSampleDynamicFrame(num:Int, options: JsonOptions = JsonOptions.empty):
 DynamicFrame

  def glueContext : GlueContext

  def setFormat( format : String,
                 options : String
               ) : Unit

  def setFormat( format : String,
                 options : JsonOptions
               ) : Unit

  def supportsFormat( format : String ) : Boolean

  def withFormat( format : String,
                  options : JsonOptions = JsonOptions.empty
                ) : DataSource
}
```

AWS Glue Scala DynamicFrame APIs

Package: com.amazonaws.services.glue

Contents

- AWS Glue Scala DynamicFrame class
  - Val errorsCount
  - Def applyMapping
  - Def assertErrorThreshold
  - Def count
  - Def dropField
  - Def dropFields
  - Def dropNulls
  - Def errorsAsDynamicFrame
  - Def filter
  - Def getName
  - Def getNumPartitions

  - Def getSchemaIfComputed
  - Def isSchemaComputed
  - Def javaToPython
  - Def join
  - Def map
  - Def mergeDynamicFrames
  - Def printSchema
  - Def recomputeSchema
  - Def relationalize
  - Def renameField
  - Def repartition
  - Def resolveChoice
  - Def schema
  - Def selectField
  - Def selectFields
  - Def show
  - Def simplifyDDBJson
  - Def spigot
  - Def splitFields
  - Def splitRows
  - Def stageErrorsCount
  - Def toDF
  - Def unbox
  - Def unnest
  - Def unnestDDBJson
  - Def withFrameSchema
  - Def withName
  - Def withTransformationContext
- The DynamicFrame object
  - Def apply

- Def emptyDynamicFrame
- Def fromPythonRDD
- Def ignoreErrors
- Def inlineErrors
- Def newFrameWithErrors

AWS Glue Scala DynamicFrame class

Package: com.amazonaws.services.glue

```
class DynamicFrame extends Serializable with Logging  (
           val glueContext : GlueContext,
           _records : RDD[DynamicRecord],
           val name : String = s"",
           val transformationContext : String = DynamicFrame.UNDEFINED,
           callSite : CallSite = CallSite("Not provided", ""),
           stageThreshold : Long = 0,
           totalThreshold : Long = 0,
           prevErrors : => Long = 0,
           errorExpr : => Unit = {} )
```

A `DynamicFrame` is a distributed collection of self-describing DynamicRecord objects.

`DynamicFrames` are designed to provide a flexible data model for ETL (extract, transform, and load) operations. They don't require a schema to create, and you can use them to read and transform data that contains messy or inconsistent values and types. A schema can be computed on demand for those operations that need one.

`DynamicFrames` provide a range of transformations for data cleaning and ETL. They also support conversion to and from SparkSQL DataFrames to integrate with existing code and the many analytics operations that DataFrames provide.

The following parameters are shared across many of the AWS Glue transformations that construct `DynamicFrames`:

- `transformationContext` — The identifier for this `DynamicFrame`. The `transformationContext` is used as a key for job bookmark state that is persisted across runs.
- `callSite` — Provides context information for error reporting. These values are automatically set when calling from Python.

- `stageThreshold` — The maximum number of error records that are allowed from the computation of this `DynamicFrame` before throwing an exception, excluding records that are present in the previous `DynamicFrame`.
- `totalThreshold` — The maximum number of total error records before an exception is thrown, including those from previous frames.

Val errorsCount

```
val errorsCount
```

The number of error records in this `DynamicFrame`. This includes errors from previous operations.

Def applyMapping

```
def applyMapping( mappings : Seq[Product4[String, String, String, String]],
                  caseSensitive : Boolean = true,
                  transformationContext : String = "",
                  callSite : CallSite = CallSite("Not provided", ""),
                  stageThreshold : Long = 0,
                  totalThreshold : Long = 0
                ) : DynamicFrame
```

- `mappings` — A sequence of mappings to construct a new `DynamicFrame`.
- `caseSensitive` — Whether to treat source columns as case sensitive. Setting this to false might help when integrating with case-insensitive stores like the AWS Glue Data Catalog.

Selects, projects, and casts columns based on a sequence of mappings.

Each mapping is made up of a source column and type and a target column and type. Mappings can be specified as either a four-tuple (`source_path`, `source_type`, `target_path`, `target_type`) or a MappingSpec object containing the same information.

In addition to using mappings for simple projections and casting, you can use them to nest or unnest fields by separating components of the path with '`.`' (period).

For example, suppose that you have a `DynamicFrame` with the following schema.

```
 {{{
  root
  |-- name: string
  |-- age: int
  |-- address: struct
  |    |-- state: string
  |    |-- zip: int
}}}
```

You can make the following call to unnest the `state` and `zip` fields.

```
{{{
  df.applyMapping(
   Seq(("name", "string", "name", "string"),
       ("age", "int", "age", "int"),
       ("address.state", "string", "state", "string"),
       ("address.zip", "int", "zip", "int")))
}}}
```

The resulting schema is as follows.

```
{{{
  root
  |-- name: string
  |-- age: int
  |-- state: string
  |-- zip: int
}}}
```

You can also use `applyMapping` to re-nest columns. For example, the following inverts the previous transformation and creates a struct named `address` in the target.

```
{{{
  df.applyMapping(
   Seq(("name", "string", "name", "string"),
       ("age", "int", "age", "int"),
       ("state", "string", "address.state", "string"),
       ("zip", "int", "address.zip", "int")))
}}}
```

Field names that contain '`.`' (period) characters can be quoted by using backticks (` `` `).

Note

Currently, you can't use the `applyMapping` method to map columns that are nested under arrays.

Def assertErrorThreshold

```
def assertErrorThreshold : Unit
```

An action that forces computation and verifies that the number of error records falls below `stageThreshold` and `totalThreshold`. Throws an exception if either condition fails.

Def count

```
lazy
def count
```

Returns the number of elements in this `DynamicFrame`.

Def dropField

```
def dropField( path : String,
               transformationContext : String = "",
               callSite : CallSite = CallSite("Not provided", ""),
               stageThreshold : Long = 0,
               totalThreshold : Long = 0
             ) : DynamicFrame
```

Returns a new `DynamicFrame` with the specified column removed.

Def dropFields

```
def dropFields( fieldNames : Seq[String],   // The column names to drop.
                transformationContext : String = "",
                callSite : CallSite = CallSite("Not provided", ""),
                stageThreshold : Long = 0,
                totalThreshold : Long = 0
              ) : DynamicFrame
```

Returns a new `DynamicFrame` with the specified columns removed.

You can use this method to delete nested columns, including those inside of arrays, but not to drop specific array elements.

Def dropNulls

```
def dropNulls( transformationContext : String = "",
               callSite : CallSite = CallSite("Not provided", ""),
               stageThreshold : Long = 0,
               totalThreshold : Long = 0 )
```

Returns a new `DynamicFrame` with all null columns removed.

> **Note**
>
> This only removes columns of type `NullType`. Individual null values in other columns are not removed or modified.

Def errorsAsDynamicFrame

```
def errorsAsDynamicFrame
```

Returns a new `DynamicFrame` containing the error records from this `DynamicFrame`.

Def filter

```
def filter( f : DynamicRecord => Boolean,
            errorMsg : String = "",
            transformationContext : String = "",
            callSite : CallSite = CallSite("Not provided"),
            stageThreshold : Long = 0,
            totalThreshold : Long = 0
          ) : DynamicFrame
```

Constructs a new `DynamicFrame` containing only those records for which the function 'f' returns `true`. The filter function 'f' should not mutate the input record.

Def getName

```
def getName : String
```

Returns the name of this `DynamicFrame`.

Def getNumPartitions

```
def getNumPartitions
```

Returns the number of partitions in this `DynamicFrame`.

Def getSchemaIfComputed

```
def getSchemaIfComputed : Option[Schema]
```

Returns the schema if it has already been computed. Does not scan the data if the schema has not already been computed.

Def isSchemaComputed

```
def isSchemaComputed : Boolean
```

Returns `true` if the schema has been computed for this `DynamicFrame`, or `false` if not. If this method returns false, then calling the `schema` method requires another pass over the records in this `DynamicFrame`.

Def javaToPython

```
def javaToPython : JavaRDD[Array[Byte]]
```

Def join

```
def join( keys1 : Seq[String],
          keys2 : Seq[String],
          frame2 : DynamicFrame,
          transformationContext : String = "",
          callSite : CallSite = CallSite("Not provided", ""),
          stageThreshold : Long = 0,
          totalThreshold : Long = 0
        ) : DynamicFrame
```

- `keys1` — The columns in this `DynamicFrame` to use for the join.

- `keys2` — The columns in `frame2` to use for the join. Must be the same length as `keys1`.
- `frame2` — The `DynamicFrame` to join against.

Returns the result of performing an equijoin with `frame2` using the specified keys.

Def map

```
def map( f : DynamicRecord => DynamicRecord,
         errorMsg : String = "",
         transformationContext : String = "",
         callSite : CallSite = CallSite("Not provided", ""),
         stageThreshold : Long = 0,
         totalThreshold : Long = 0
       ) : DynamicFrame
```

Returns a new `DynamicFrame` constructed by applying the specified function 'f' to each record in this `DynamicFrame`.

This method copies each record before applying the specified function, so it is safe to mutate the records. If the mapping function throws an exception on a given record, that record is marked as an error, and the stack trace is saved as a column in the error record.

Def mergeDynamicFrames

```
def mergeDynamicFrames( stageDynamicFrame: DynamicFrame,  primaryKeys: Seq[String],
 transformationContext: String = "",
                         options: JsonOptions = JsonOptions.empty, callSite: CallSite =
 CallSite("Not provided"),
                         stageThreshold: Long = 0, totalThreshold: Long = 0):
 DynamicFrame
```

- `stageDynamicFrame` — The staging `DynamicFrame` to merge.
- `primaryKeys` — The list of primary key fields to match records from the source and staging `DynamicFrames`.
- `transformationContext` — A unique string that is used to retrieve metadata about the current transformation (optional).
- `options` — A string of JSON name-value pairs that provide additional information for this transformation.

- `callSite` — Used to provide context information for error reporting.
- `stageThreshold` — A `Long`. The number of errors in the given transformation for which the processing needs to error out.
- `totalThreshold` — A `Long`. The total number of errors up to and including in this transformation for which the processing needs to error out.

Merges this `DynamicFrame` with a staging `DynamicFrame` based on the specified primary keys to identify records. Duplicate records (records with the same primary keys) are not de-duplicated. If there is no matching record in the staging frame, all records (including duplicates) are retained from the source. If the staging frame has matching records, the records from the staging frame overwrite the records in the source in AWS Glue.

The returned `DynamicFrame` contains record A in the following cases:

1. If `A` exists in both the source frame and the staging frame, then `A` in the staging frame is returned.
2. If `A` is in the source table and `A.primaryKeys` is not in the `stagingDynamicFrame` (that means `A` is not updated in the staging table).

The source frame and staging frame do not need to have the same schema.

Example

```
val mergedFrame: DynamicFrame = srcFrame.mergeDynamicFrames(stageFrame, Seq("id1",
 "id2"))
```

Def printSchema

```
def printSchema : Unit
```

Prints the schema of this `DynamicFrame` to `stdout` in a human-readable format.

Def recomputeSchema

```
def recomputeSchema : Schema
```

Forces a schema recomputation. This requires a scan over the data, but it might "tighten" the schema if there are some fields in the current schema that are not present in the data.

Returns the recomputed schema.

Def relationalize

```
def relationalize( rootTableName : String,
                   stagingPath : String,
                   options : JsonOptions = JsonOptions.empty,
                   transformationContext : String = "",
                   callSite : CallSite = CallSite("Not provided"),
                   stageThreshold : Long = 0,
                   totalThreshold : Long = 0
                 ) : Seq[DynamicFrame]
```

- `rootTableName` — The name to use for the base `DynamicFrame` in the output. `DynamicFrames` that are created by pivoting arrays start with this as a prefix.
- `stagingPath` — The Amazon Simple Storage Service (Amazon S3) path for writing intermediate data.
- `options` — Relationalize options and configuration. Currently unused.

Flattens all nested structures and pivots arrays into separate tables.

You can use this operation to prepare deeply nested data for ingestion into a relational database. Nested structs are flattened in the same manner as the Unnest transform. Additionally, arrays are pivoted into separate tables with each array element becoming a row. For example, suppose that you have a `DynamicFrame` with the following data.

```
 {"name": "Nancy", "age": 47, "friends": ["Fred", "Lakshmi"]}
 {"name": "Stephanie", "age": 28, "friends": ["Yao", "Phil", "Alvin"]}
 {"name": "Nathan", "age": 54, "friends": ["Nicolai", "Karen"]}
```

Run the following code.

```
{{{
  df.relationalize("people", "s3:/my_bucket/my_path", JsonOptions.empty)
}}}
```

This produces two tables. The first table is named "people" and contains the following.

```
{{{
  {"name": "Nancy", "age": 47, "friends": 1}
  {"name": "Stephanie", "age": 28, "friends": 2}
  {"name": "Nathan", "age": 54, "friends": 3)
}}}
```

Here, the friends array has been replaced with an auto-generated join key. A separate table named `people.friends` is created with the following content.

```
{{{
  {"id": 1, "index": 0, "val": "Fred"}
  {"id": 1, "index": 1, "val": "Lakshmi"}
  {"id": 2, "index": 0, "val": "Yao"}
  {"id": 2, "index": 1, "val": "Phil"}
  {"id": 2, "index": 2, "val": "Alvin"}
  {"id": 3, "index": 0, "val": "Nicolai"}
  {"id": 3, "index": 1, "val": "Karen"}
}}}
```

In this table, '`id`' is a join key that identifies which record the array element came from, '`index`' refers to the position in the original array, and '`val`' is the actual array entry.

The `relationalize` method returns the sequence of `DynamicFrames` created by applying this process recursively to all arrays.

> **Note**
>
> The AWS Glue library automatically generates join keys for new tables. To ensure that join keys are unique across job runs, you must enable job bookmarks.

Def renameField

```
def renameField( oldName : String,
                 newName : String,
                 transformationContext : String = "",
                 callSite : CallSite = CallSite("Not provided", ""),
                 stageThreshold : Long = 0,
                 totalThreshold : Long = 0
               ) : DynamicFrame
```

- `oldName` — The original name of the column.

- `newName` — The new name of the column.

Returns a new `DynamicFrame` with the specified field renamed.

You can use this method to rename nested fields. For example, the following code would rename `state` to `state_code` inside the address struct.

```
{{{
  df.renameField("address.state", "address.state_code")
}}}
```

Def repartition

```
def repartition( numPartitions : Int,
                 transformationContext : String = "",
                 callSite : CallSite = CallSite("Not provided", ""),
                 stageThreshold : Long = 0,
                 totalThreshold : Long = 0
               ) : DynamicFrame
```

Returns a new `DynamicFrame` with `numPartitions` partitions.

Def resolveChoice

```
def resolveChoice( specs : Seq[Product2[String, String]] = Seq.empty[ResolveSpec],
                   choiceOption : Option[ChoiceOption] = None,
                   database : Option[String] = None,
                   tableName : Option[String] = None,
                   transformationContext : String = "",
                   callSite : CallSite = CallSite("Not provided", ""),
                   stageThreshold : Long = 0,
                   totalThreshold : Long = 0
                 ) : DynamicFrame
```

- `choiceOption` — An action to apply to all `ChoiceType` columns not listed in the specs sequence.
- `database` — The Data Catalog database to use with the `match_catalog` action.
- `tableName` — The Data Catalog table to use with the `match_catalog` action.

Returns a new `DynamicFrame` by replacing one or more `ChoiceTypes` with a more specific type.

There are two ways to use `resolveChoice`. The first is to specify a sequence of specific columns and how to resolve them. These are specified as tuples made up of (column, action) pairs.

The following are the possible actions:

- `cast:type` — Attempts to cast all values to the specified type.
- `make_cols` — Converts each distinct type to a column with the name `columnName_type`.
- `make_struct` — Converts a column to a struct with keys for each distinct type.
- `project:type` — Retains only values of the specified type.

The other mode for `resolveChoice` is to specify a single resolution for all `ChoiceTypes`. You can use this in cases where the complete list of `ChoiceTypes` is unknown before execution. In addition to the actions listed preceding, this mode also supports the following action:

- `match_catalog` — Attempts to cast each `ChoiceType` to the corresponding type in the specified catalog table.

Examples:

Resolve the `user.id` column by casting to an int, and make the `address` field retain only structs.

```
{{{
  df.resolveChoice(specs = Seq(("user.id", "cast:int"), ("address", "project:struct")))
}}}
```

Resolve all `ChoiceTypes` by converting each choice to a separate column.

```
{{{
  df.resolveChoice(choiceOption = Some(ChoiceOption("make_cols")))
}}}
```

Resolve all `ChoiceTypes` by casting to the types in the specified catalog table.

```
{{{
  df.resolveChoice(choiceOption = Some(ChoiceOption("match_catalog")),
                   database = Some("my_database"),
                   tableName = Some("my_table"))
}}}
```

Def schema

```
def schema : Schema
```

Returns the schema of this `DynamicFrame`.

The returned schema is guaranteed to contain every field that is present in a record in this `DynamicFrame`. But in a small number of cases, it might also contain additional fields. You can use the Unnest method to "tighten" the schema based on the records in this `DynamicFrame`.

Def selectField

```
def selectField( fieldName : String,
                 transformationContext : String = "",
                 callSite : CallSite = CallSite("Not provided", ""),
                 stageThreshold : Long = 0,
                 totalThreshold : Long = 0
               ) : DynamicFrame
```

Returns a single field as a `DynamicFrame`.

Def selectFields

```
def selectFields( paths : Seq[String],
                  transformationContext : String = "",
                  callSite : CallSite = CallSite("Not provided", ""),
                  stageThreshold : Long = 0,
                  totalThreshold : Long = 0
                ) : DynamicFrame
```

- `paths` — The sequence of column names to select.

Returns a new `DynamicFrame` containing the specified columns.

> **Note**
>
> You can only use the `selectFields` method to select top-level columns. You can use the applyMapping method to select nested columns.

Def show

```
def show( numRows : Int = 20 ) : Unit
```

- `numRows` — The number of rows to print.

Prints rows from this `DynamicFrame` in JSON format.

Def simplifyDDBJson

DynamoDB exports with the AWS Glue DynamoDB export connector results in JSON files of specific nested structures. For more information, see Data objects. `simplifyDDBJson` Simplifies nested columns in a DynamicFrame of this type of data, and returns a new simplified DynamicFrame. If there are multiple types or a Map type contained in a List type, the elements in the List will not be simplified. This method only supports data in the DynamoDB export JSON format. Consider `unnest` to perform similar changes on other kinds of data.

```
def simplifyDDBJson() : DynamicFrame
```

This method does not take any parameters.

Example input

Consider the following schema generated by a DynamoDB export:

```
root
|-- Item: struct
|    |-- parentMap: struct
|    |    |-- M: struct
|    |    |    |-- childMap: struct
|    |    |    |    |-- M: struct
|    |    |    |    |    |-- appName: struct
|    |    |    |    |    |    |-- S: string
|    |    |    |    |    |-- packageName: struct
|    |    |    |    |    |    |-- S: string
|    |    |    |    |    |-- updatedAt: struct
|    |    |    |    |    |    |-- N: string
|    |-- strings: struct
|    |    |-- SS: array
|    |    |    |-- element: string
|    |-- numbers: struct
|     |     |-- NS: array
|     |     |     |-- element: string
|     |-- binaries: struct
|     |     |-- BS: array
|     |     |     |-- element: string
|     |-- isDDBJson: struct
|     |     |-- BOOL: boolean
|     |-- nullValue: struct
|     |     |-- NULL: boolean
```

Example code

```
import com.amazonaws.services.glue.GlueContext
import com.amazonaws.services.glue.util.GlueArgParser
import com.amazonaws.services.glue.util.Job
import com.amazonaws.services.glue.util.JsonOptions
import com.amazonaws.services.glue.DynamoDbDataSink
import org.apache.spark.SparkContextimport scala.collection.JavaConverters._

object GlueApp {

  def main(sysArgs: Array[String]): Unit = {
    val glueContext = new GlueContext(SparkContext.getOrCreate())
    val args = GlueArgParser.getResolvedOptions(sysArgs, Seq("JOB_NAME").toArray)
    Job.init(args("JOB_NAME"), glueContext, args.asJava)

    val dynamicFrame = glueContext.getSourceWithFormat(
      connectionType = "dynamodb",
      options = JsonOptions(Map(
        "dynamodb.export" -> "ddb",
        "dynamodb.tableArn" -> "ddbTableARN",
        "dynamodb.s3.bucket" -> "exportBucketLocation",
        "dynamodb.s3.prefix" -> "exportBucketPrefix",
        "dynamodb.s3.bucketOwner" -> "exportBucketAccountID",
      ))
    ).getDynamicFrame()

    val simplified = dynamicFrame.simplifyDDBJson()
    simplified.printSchema()

    Job.commit()
  }
}
```

Example output

The `simplifyDDBJson` transform will simplify this to:

```
root
|-- parentMap: struct
|    |-- childMap: struct
|    |    |-- appName: string
|    |    |-- packageName: string
|    |    |-- updatedAt: string
|-- strings: array
|    |-- element: string
|-- numbers: array
|    |-- element: string
|-- binaries: array
|    |-- element: string
|-- isDDBJson: boolean
|-- nullValue: null
```

Def spigot

```
def spigot( path : String,
            options : JsonOptions = new JsonOptions("{}"),
            transformationContext : String = "",
            callSite : CallSite = CallSite("Not provided"),
            stageThreshold : Long = 0,
            totalThreshold : Long = 0
          ) : DynamicFrame
```

Passthrough transformation that returns the same records but writes out a subset of records as a side effect.

- `path` — The path in Amazon S3 to write output to, in the form `s3://bucket//path`.
- `options` — An optional `JsonOptions` map describing the sampling behavior.

Returns a `DynamicFrame` that contains the same records as this one.

By default, writes 100 arbitrary records to the location specified by `path`. You can customize this behavior by using the `options` map. Valid keys include the following:

- `topk` — Specifies the total number of records written out. The default is 100.
- `prob` — Specifies the probability (as a decimal) that an individual record is included. Default is 1.

For example, the following call would sample the dataset by selecting each record with a 20 percent probability and stopping after 200 records have been written.

```
{{{
  df.spigot("s3://my_bucket/my_path", JsonOptions(Map("topk" -&gt; 200, "prob" -&gt;
 0.2)))
}}}
```

Def splitFields

```
def splitFields( paths : Seq[String],
                 transformationContext : String = "",
                 callSite : CallSite = CallSite("Not provided", ""),
                 stageThreshold : Long = 0,
                 totalThreshold : Long = 0
               ) : Seq[DynamicFrame]
```

- `paths` — The paths to include in the first `DynamicFrame`.

Returns a sequence of two `DynamicFrames`. The first `DynamicFrame` contains the specified paths, and the second contains all other columns.

Example

This example takes a DynamicFrame created from the `persons` table in the `legislators` database in the AWS Glue Data Catalog and splits the DynamicFrame into two, with the specified fields going into the first DynamicFrame and the remaining fields going into a second DynamicFrame. The example then chooses the first DynamicFrame from the result.

```
val InputFrame = glueContext.getCatalogSource(database="legislators",
 tableName="persons",
transformationContext="InputFrame").getDynamicFrame()

val SplitField_collection = InputFrame.splitFields(paths=Seq("family_name", "name",
 "links.note",
"links.url", "gender", "image", "identifiers.scheme", "identifiers.identifier",
 "other_names.lang",
"other_names.note", "other_names.name"), transformationContext="SplitField_collection")

val ResultFrame = SplitField_collection(0)
```

Def splitRows

```
def splitRows( paths : Seq[String],
               values : Seq[Any],
               operators : Seq[String],
               transformationContext : String,
               callSite : CallSite,
               stageThreshold : Long,
               totalThreshold : Long
             ) : Seq[DynamicFrame]
```

Splits rows based on predicates that compare columns to constants.

- `paths` — The columns to use for comparison.
- `values` — The constant values to use for comparison.
- `operators` — The operators to use for comparison.

Returns a sequence of two `DynamicFrames`. The first contains rows for which the predicate is true and the second contains those for which it is false.

Predicates are specified using three sequences: '`paths`' contains the (possibly nested) column names, '`values`' contains the constant values to compare to, and '`operators`' contains the operators to use for comparison. All three sequences must be the same length: The `n`th operator is used to compare the `n`th column with the `n`th value.

Each operator must be one of "`!=`", "`=`", "`<=`", "`<`", "`>=`", or "`>`".

As an example, the following call would split a `DynamicFrame` so that the first output frame would contain records of people over 65 from the United States, and the second would contain all other records.

```
{{{
  df.splitRows(Seq("age", "address.country"), Seq(65, "USA"), Seq("&gt;=", "="))
}}}
```

Def stageErrorsCount

```
def stageErrorsCount
```

Returns the number of error records created while computing this `DynamicFrame`. This excludes errors from previous operations that were passed into this `DynamicFrame` as input.

Def toDF

```
def toDF( specs : Seq[ResolveSpec] = Seq.empty[ResolveSpec] ) : DataFrame
```

Converts this `DynamicFrame` to an Apache Spark SQL `DataFrame` with the same schema and records.

> **Note**
>
> Because `DataFrames` don't support `ChoiceTypes`, this method automatically converts `ChoiceType` columns into `StructTypes`. For more information and options for resolving choice, see resolveChoice.

Def unbox

```
def unbox( path : String,
           format : String,
           optionString : String = "{}",
           transformationContext : String = "",
           callSite : CallSite = CallSite("Not provided"),
           stageThreshold : Long = 0,
           totalThreshold : Long = 0
         ) : DynamicFrame
```

- `path` — The column to parse. Must be a string or binary.
- `format` — The format to use for parsing.
- `optionString` — Options to pass to the format, such as the CSV separator.

Parses an embedded string or binary column according to the specified format. Parsed columns are nested under a struct with the original column name.

For example, suppose that you have a CSV file with an embedded JSON column.

```
name, age, address
Sally, 36, {"state": "NE", "city": "Omaha"}
...
```

After an initial parse, you would get a `DynamicFrame` with the following schema.

```
{{{
  root
  |-- name: string
  |-- age: int
  |-- address: string
}}}
```

You can call `unbox` on the address column to parse the specific components.

```
{{{
  df.unbox("address", "json")
}}}
```

This gives us a `DynamicFrame` with the following schema.

```
{{{
  root
  |-- name: string
  |-- age: int
  |-- address: struct
  |    |-- state: string
  |    |-- city: string
}}}
```

Def unnest

```
def unnest( transformationContext : String = "",
            callSite : CallSite = CallSite("Not Provided"),
            stageThreshold : Long = 0,
            totalThreshold : Long = 0
          ) : DynamicFrame
```

Returns a new `DynamicFrame` with all nested structures flattened. Names are constructed using the `'.'` (period) character.

For example, suppose that you have a `DynamicFrame` with the following schema.

```
{{{
  root
  |-- name: string
  |-- age: int
  |-- address: struct
  |    |-- state: string
  |    |-- city: string
}}}
```

The following call unnests the address struct.

```
{{{
  df.unnest()
}}}
```

The resulting schema is as follows.

```
{{{
  root
  |-- name: string
  |-- age: int
  |-- address.state: string
  |-- address.city: string
}}}
```

This method also unnests nested structs inside of arrays. But for historical reasons, the names of such fields are prepended with the name of the enclosing array and "`.val`".

Def unnestDDBJson

```
unnestDDBJson(transformationContext : String = "",
            callSite : CallSite = CallSite("Not Provided"),
            stageThreshold : Long = 0,
            totalThreshold : Long = 0): DynamicFrame
```

Unnests nested columns in a `DynamicFrame` that are specifically in the DynamoDB JSON structure, and returns a new unnested `DynamicFrame`. Columns that are of an array of struct types will not be unnested. Note that this is a specific type of unnesting transform that behaves differently from the regular `unnest` transform and requires the data to already be in the DynamoDB JSON structure. For more information, see DynamoDB JSON.

For example, the schema of a reading an export with the DynamoDB JSON structure might look like the following:

```
root
|-- Item: struct
|    |-- ColA: struct
|    |    |-- S: string
|    |-- ColB: struct
|    |    |-- S: string
|    |-- ColC: struct
|    |    |-- N: string
|    |-- ColD: struct
|    |    |-- L: array
|    |    |    |-- element: null
```

The `unnestDDBJson()` transform would convert this to:

```
root
|-- ColA: string
|-- ColB: string
|-- ColC: string
|-- ColD: array
|    |-- element: null
```

The following code example shows how to use the AWS Glue DynamoDB export connector, invoke a DynamoDB JSON unnest, and print the number of partitions:

```
import com.amazonaws.services.glue.GlueContext
import com.amazonaws.services.glue.util.GlueArgParser
import com.amazonaws.services.glue.util.Job
import com.amazonaws.services.glue.util.JsonOptions
import com.amazonaws.services.glue.DynamoDbDataSink
import org.apache.spark.SparkContext
import scala.collection.JavaConverters._
```

```
object GlueApp {

  def main(sysArgs: Array[String]): Unit = {
    val glueContext = new GlueContext(SparkContext.getOrCreate())
    val args = GlueArgParser.getResolvedOptions(sysArgs, Seq("JOB_NAME").toArray)
    Job.init(args("JOB_NAME"), glueContext, args.asJava)

    val dynamicFrame = glueContext.getSourceWithFormat(
      connectionType = "dynamodb",
      options = JsonOptions(Map(
        "dynamodb.export" -> "ddb",
        "dynamodb.tableArn" -> "<test_source>",
        "dynamodb.s3.bucket" -> "<bucket name>",
        "dynamodb.s3.prefix" -> "<bucket prefix>",
        "dynamodb.s3.bucketOwner" -> "<account_id of bucket>",
      ))
    ).getDynamicFrame()

    val unnested = dynamicFrame.unnestDDBJson()
    print(unnested.getNumPartitions())

    Job.commit()
  }

}
```

Def withFrameSchema

```
def withFrameSchema( getSchema : () => Schema ) : DynamicFrame
```

- `getSchema` — A function that returns the schema to use. Specified as a zero-parameter function to defer potentially expensive computation.

Sets the schema of this `DynamicFrame` to the specified value. This is primarily used internally to avoid costly schema recomputation. The passed-in schema must contain all columns present in the data.

Def withName

```
def withName( name : String ) : DynamicFrame
```

- name — The new name to use.

Returns a copy of this `DynamicFrame` with a new name.

Def withTransformationContext

```
def withTransformationContext( ctx : String ) : DynamicFrame
```

Returns a copy of this `DynamicFrame` with the specified transformation context.

The DynamicFrame object

Package: com.amazonaws.services.glue

```
object DynamicFrame
```

Def apply

```
def apply( df : DataFrame,
           glueContext : GlueContext
         ) : DynamicFrame
```

Def emptyDynamicFrame

```
def emptyDynamicFrame( glueContext : GlueContext ) : DynamicFrame
```

Def fromPythonRDD

```
def fromPythonRDD( rdd : JavaRDD[Array[Byte]],
                   glueContext : GlueContext
                 ) : DynamicFrame
```

Def ignoreErrors

```
def ignoreErrors( fn : DynamicRecord => DynamicRecord ) : DynamicRecord
```

Def inlineErrors

```
def inlineErrors( msg : String,
                  callSite : CallSite
                ) : (DynamicRecord => DynamicRecord)
```

Def newFrameWithErrors

```
def newFrameWithErrors( prevFrame : DynamicFrame,
                        rdd : RDD[DynamicRecord],
                        name : String = "",
                        transformationContext : String = "",
                        callSite : CallSite,
                        stageThreshold : Long,
                        totalThreshold : Long
                      ) : DynamicFrame
```

AWS Glue Scala DynamicRecord class

Topics

- Def addField
- Def dropField
- Def setError
- Def isError
- Def getError
- Def clearError
- Def write
- Def readFields
- Def clone
- Def schema
- Def getRoot
- Def toJson
- Def getFieldNode
- Def getField
- Def hashCode

- Def equals
- DynamicRecord object
- RecordTraverser trait

Package: com.amazonaws.services.glue

```
class DynamicRecord extends Serializable with Writable with Cloneable
```

A `DynamicRecord` is a self-describing data structure that represents a row of data in the dataset that is being processed. It is self-describing in the sense that you can get the schema of the row that is represented by the `DynamicRecord` by inspecting the record itself. A `DynamicRecord` is similar to a Row in Apache Spark.

Def addField

```
def addField( path : String,
              dynamicNode : DynamicNode
            ) : Unit
```

Adds a DynamicNode to the specified path.

- `path` — The path for the field to be added.
- `dynamicNode` — The DynamicNode to be added at the specified path.

Def dropField

```
 def dropField(path: String, underRename: Boolean = false): Option[DynamicNode]
```

Drops a DynamicNode from the specified path and returns the dropped node if there is not an array in the specified path.

- `path` — The path to the field to drop.
- `underRename` — True if `dropField` is called as part of a rename transform, or false otherwise (false by default).

Returns a `scala.Option Option` (DynamicNode).

Def setError

```
def setError( error : Error )
```

Sets this record as an error record, as specified by the `error` parameter.

Returns a `DynamicRecord`.

Def isError

```
def isError
```

Checks whether this record is an error record.

Def getError

```
def getError
```

Gets the `Error` if the record is an error record. Returns `scala.Some Some` (Error) if this record is an error record, or otherwise `scala.None` .

Def clearError

```
def clearError
```

Set the `Error` to `scala.None.None` .

Def write

```
override def write( out : DataOutput ) : Unit
```

Def readFields

```
override def readFields( in : DataInput ) : Unit
```

Def clone

```
override def clone : DynamicRecord
```

Clones this record to a new `DynamicRecord` and returns it.

Def schema

```
def schema
```

Gets the `Schema` by inspecting the record.

Def getRoot

```
def getRoot : ObjectNode
```

Gets the root `ObjectNode` for the record.

Def toJson

```
def toJson : String
```

Gets the JSON string for the record.

Def getFieldNode

```
def getFieldNode( path : String ) : Option[DynamicNode]
```

Gets the field's value at the specified `path` as an option of `DynamicNode`.

Returns `scala.Some Some` (DynamicNode) if the field exists, or otherwise `scala.None.None` .

Def getField

```
def getField( path : String ) : Option[Any]
```

Gets the field's value at the specified `path` as an option of `DynamicNode`.

Returns `scala.Some Some` (value).

Def hashCode

```
override def hashCode : Int
```

Def equals

```
override def equals( other : Any )
```

DynamicRecord object

```
object DynamicRecord
```

Def apply

```
def apply( row : Row,
           schema : SparkStructType )
```

Apply method to convert an Apache Spark SQL Row to a DynamicRecord.

- `row` — A Spark SQL Row.
- `schema` — The Schema of that row.

Returns a `DynamicRecord`.

RecordTraverser trait

```
trait RecordTraverser {
  def nullValue(): Unit
  def byteValue(value: Byte): Unit
  def binaryValue(value: Array[Byte]): Unit
  def booleanValue(value: Boolean): Unit
  def shortValue(value: Short) : Unit
  def intValue(value: Int) : Unit
  def longValue(value: Long) : Unit
  def floatValue(value: Float): Unit
  def doubleValue(value: Double): Unit
  def decimalValue(value: BigDecimal): Unit
  def stringValue(value: String): Unit
  def dateValue(value: Date): Unit
  def timestampValue(value: Timestamp): Unit
  def objectStart(length: Int): Unit
  def objectKey(key: String): Unit
  def objectEnd(): Unit
  def mapStart(length: Int): Unit
  def mapKey(key: String): Unit
  def mapEnd(): Unit
  def arrayStart(length: Int): Unit
  def arrayEnd(): Unit
}
```

AWS Glue Scala GlueContext APIs

Package: com.amazonaws.services.glue

```
class GlueContext extends SQLContext(sc) (
           @transient val sc : SparkContext,
           val defaultSourcePartitioner : PartitioningStrategy )
```

`GlueContext` is the entry point for reading and writing a DynamicFrame from and to Amazon Simple Storage Service (Amazon S3), the AWS Glue Data Catalog, JDBC, and so on. This class provides utility functions to create DataSource trait and DataSink objects that can in turn be used to read and write `DynamicFrames`.

You can also use `GlueContext` to set a target number of partitions (default 20) in the `DynamicFrame` if the number of partitions created from the source is less than a minimum threshold for partitions (default 10).

def addIngestionTimeColumns

```
def addIngestionTimeColumns(
         df : DataFrame,
         timeGranularity : String = "") : dataFrame
```

Appends ingestion time columns like `ingest_year`, `ingest_month`, `ingest_day`, `ingest_hour`, `ingest_minute` to the input `DataFrame`. This function is automatically generated in the script generated by the AWS Glue when you specify a Data Catalog table with Amazon S3 as the target. This function automatically updates the partition with ingestion time columns on the output table. This allows the output data to be automatically partitioned on ingestion time without requiring explicit ingestion time columns in the input data.

- `dataFrame` – The `dataFrame` to append the ingestion time columns to.
- `timeGranularity` – The granularity of the time columns. Valid values are "`day`", "`hour`" and "`minute`". For example, if "`hour`" is passed in to the function, the original `dataFrame` will have "`ingest_year`", "`ingest_month`", "`ingest_day`", and "`ingest_hour`" time columns appended.

Returns the data frame after appending the time granularity columns.

Example:

```
glueContext.addIngestionTimeColumns(dataFrame, "hour")
```

def createDataFrameFromOptions

```
def createDataFrameFromOptions( connectionType : String,
                         connectionOptions : JsonOptions,
                         transformationContext : String = "",
                         format : String = null,
                         formatOptions : JsonOptions = JsonOptions.empty
                       ) : DataSource
```

Returns a `DataFrame` created with the specified connection and format. Use this function only with AWS Glue streaming sources.

- `connectionType` – The streaming connection type. Valid values include `kinesis` and `kafka`.
- `connectionOptions` – Connection options, which are different for Kinesis and Kafka. You can find the list of all connection options for each streaming data source at Connection types and options for ETL in AWS Glue for Spark. Note the following differences in streaming connection options:
  - Kinesis streaming sources require `streamARN`, `startingPosition`, `inferSchema`, and `classification`.
  - Kafka streaming sources require `connectionName`, `topicName`, `startingOffsets`, `inferSchema`, and `classification`.
- `transformationContext` – The transformation context to use (optional).
- `format` – A format specification (optional). This is used for an Amazon S3 or an AWS Glue connection that supports multiple formats. For information about the supported formats, see Data format options for inputs and outputs in AWS Glue for Spark
- `formatOptions` – Format options for the specified format. For information about the supported format options, see Data format options.

Example for Amazon Kinesis streaming source:

```
val data_frame_datasource0 =
glueContext.createDataFrameFromOptions(transformationContext = "datasource0",
 connectionType = "kinesis",
connectionOptions = JsonOptions("""{"streamName": "example_stream", "startingPosition":
 "TRIM_HORIZON", "inferSchema": "true", "classification": "json"}}"""))
```

Example for Kafka streaming source:

```
val data_frame_datasource0 =
glueContext.createDataFrameFromOptions(transformationContext = "datasource0",
 connectionType = "kafka",
connectionOptions = JsonOptions("""{"connectionName": "example_connection",
 "topicName": "example_topic", "startingPosition": "earliest", "inferSchema": "false",
 "classification": "json", "schema":"`column1` STRING, `column2` STRING"}"""))
```

forEachBatch

forEachBatch(frame, batch_function, options)

Applies the `batch_function` passed in to every micro batch that is read from the Streaming source.

- `frame` – The DataFrame containing the current micro batch.
- `batch_function` – A function that will be applied for every micro batch.
- `options` – A collection of key-value pairs that holds information about how to process micro batches. The following options are required:
  - `windowSize` – The amount of time to spend processing each batch.
  - `checkpointLocation` – The location where checkpoints are stored for the streaming ETL job.
  - `batchMaxRetries` – The maximum number of times to retry the batch if it fails. The default value is 3. This option is only configurable for Glue version 2.0 and above.

Example:

```
glueContext.forEachBatch(data_frame_datasource0, (dataFrame: Dataset[Row], batchId:
 Long) =>
   {
      if (dataFrame.count() > 0)
        {
          val datasource0 = DynamicFrame(glueContext.addIngestionTimeColumns(dataFrame,
 "hour"), glueContext)
          // @type: DataSink
          // @args: [database = "tempdb", table_name = "fromoptionsoutput",
 stream_batch_time = "100 seconds",
          //      stream_checkpoint_location = "s3://from-options-testing-eu-central-1/
fromOptionsOutput/checkpoint/",
          //      transformation_ctx = "datasink1"]
          // @return: datasink1
          // @inputs: [frame = datasource0]
          val options_datasink1 = JsonOptions(
             Map("partitionKeys" -> Seq("ingest_year", "ingest_month","ingest_day",
 "ingest_hour"),
             "enableUpdateCatalog" -> true))
          val datasink1 = glueContext.getCatalogSink(
             database = "tempdb",
             tableName = "fromoptionsoutput",
             redshiftTmpDir = "",
             transformationContext = "datasink1",
             additionalOptions = options_datasink1).writeDynamicFrame(datasource0)
        }
   }, JsonOptions("""{"windowSize" : "100 seconds",
         "checkpointLocation" : "s3://from-options-testing-eu-central-1/
fromOptionsOutput/checkpoint/"}"""))
```

def getCatalogSink

```
def getCatalogSink( database : String,
        tableName : String,
        redshiftTmpDir : String = "",
        transformationContext : String = ""
        additionalOptions: JsonOptions = JsonOptions.empty,
        catalogId: String = null
) : DataSink
```

Creates a DataSink that writes to a location specified in a table that is defined in the Data Catalog.

- `database` — The database name in the Data Catalog.
- `tableName` — The table name in the Data Catalog.
- `redshiftTmpDir` — The temporary staging directory to be used with certain data sinks. Set to empty by default.
- `transformationContext` — The transformation context that is associated with the sink to be used by job bookmarks. Set to empty by default.
- `additionalOptions` – Additional options provided to AWS Glue.

- `catalogId` — The catalog ID (account ID) of the Data Catalog being accessed. When null, the default account ID of the caller is used.

Returns the `DataSink`.

def getCatalogSource

```
def getCatalogSource( database : String,
                      tableName : String,
                      redshiftTmpDir : String = "",
                      transformationContext : String = ""
                      pushDownPredicate : String = " "
                      additionalOptions: JsonOptions = JsonOptions.empty,
                      catalogId: String = null
                    ) : DataSource
```

Creates a DataSource trait that reads data from a table definition in the Data Catalog.

- `database` — The database name in the Data Catalog.
- `tableName` — The table name in the Data Catalog.
- `redshiftTmpDir` — The temporary staging directory to be used with certain data sinks. Set to empty by default.
- `transformationContext` — The transformation context that is associated with the sink to be used by job bookmarks. Set to empty by default.
- `pushDownPredicate` – Filters partitions without having to list and read all the files in your dataset. For more information, see Pre-filtering using pushdown predicates.
- `additionalOptions` – A collection of optional name-value pairs. The possible options include those listed in Connection types and options for ETL in AWS Glue for Spark except for `endpointUrl`, `streamName`, `bootstrap.servers`, `security.protocol`, `topicName`, `classification`, and `delimiter`. Another supported option is `catalogPartitionPredicate`:

  `catalogPartitionPredicate` — You can pass a catalog expression to filter based on the index columns. This pushes down the filtering to the server side. For more information, see AWS Glue Partition Indexes. Note that `push_down_predicate` and `catalogPartitionPredicate` use different syntaxes. The former one uses Spark SQL standard syntax and the later one uses JSQL parser.

- `catalogId` — The catalog ID (account ID) of the Data Catalog being accessed. When null, the default account ID of the caller is used.

Returns the `DataSource`.

Example for streaming source

```
val data_frame_datasource0 = glueContext.getCatalogSource(
    database = "tempdb",
    tableName = "test-stream-input",
    redshiftTmpDir = "",
    transformationContext = "datasource0",
    additionalOptions = JsonOptions("""{
        "startingPosition": "TRIM_HORIZON", "inferSchema": "false"}""")
    ).getDataFrame()
```

def getJDBCSink

```
def getJDBCSink( catalogConnection : String,
                 options : JsonOptions,
                 redshiftTmpDir : String = "",
                 transformationContext : String = "",
                 catalogId: String = null
               ) : DataSink
```

Creates a DataSink that writes to a JDBC database that is specified in a `Connection` object in the Data Catalog. The `Connection` object has information to connect to a JDBC sink, including the URL, user name, password, VPC, subnet, and security groups.

- `catalogConnection` — The name of the connection in the Data Catalog that contains the JDBC URL to write to.
- `options` — A string of JSON name-value pairs that provide additional information that is required to write to a JDBC data store. This includes:
    - *dbtable* (required) — The name of the JDBC table. For JDBC data stores that support schemas within a database, specify `schema.table-name`. If a schema is not provided, then the default "public" schema is used. The following example shows an options parameter that points to a schema named `test` and a table named `test_table` in database `test_db`.

    ```
    options = JsonOptions("""{"dbtable": "test.test_table", "database": "test_db"}""")
    ```

  - *database* (required) — The name of the JDBC database.
  - Any additional options passed directly to the SparkSQL JDBC writer. For more information, see [Redshift data source for Spark](#).
- `redshiftTmpDir` — A temporary staging directory to be used with certain data sinks. Set to empty by default.
- `transformationContext` — The transformation context that is associated with the sink to be used by job bookmarks. Set to empty by default.
- `catalogId` — The catalog ID (account ID) of the Data Catalog being accessed. When null, the default account ID of the caller is used.

Example code:

```
getJDBCSink(catalogConnection = "my-connection-name", options =
 JsonOptions("""{"dbtable": "my-jdbc-table", "database": "my-jdbc-db"}"""),
 redshiftTmpDir = "", transformationContext = "datasink4")
```

Returns the `DataSink`.

def getSink

```
def getSink( connectionType : String,
             connectionOptions : JsonOptions,
             transformationContext : String = ""
           ) : DataSink
```

Creates a [DataSink](#) that writes data to a destination like Amazon Simple Storage Service (Amazon S3), JDBC, or the AWS Glue Data Catalog, or an Apache Kafka or Amazon Kinesis data stream.

- `connectionType` — The type of the connection. See [the section called “Connection parameters”](#).
- `connectionOptions` — A string of JSON name-value pairs that provide additional information to establish the connection with the data sink. See [the section called “Connection parameters”](#).
- `transformationContext` — The transformation context that is associated with the sink to be used by job bookmarks. Set to empty by default.

Returns the `DataSink`.

def getSinkWithFormat

```
def getSinkWithFormat( connectionType : String,
                       options : JsonOptions,
                       transformationContext : String = "",
                       format : String = null,
                       formatOptions : JsonOptions = JsonOptions.empty
                     ) : DataSink
```

Creates a DataSink that writes data to a destination like Amazon S3, JDBC, or the Data Catalog, or an Apache Kafka or Amazon Kinesis data stream. Also sets the format for the data to be written out to the destination.

- `connectionType` — The type of the connection. See the section called “Connection parameters”.
- `options` — A string of JSON name-value pairs that provide additional information to establish a connection with the data sink. See the section called “Connection parameters”.
- `transformationContext` — The transformation context that is associated with the sink to be used by job bookmarks. Set to empty by default.
- `format` — The format of the data to be written out to the destination.
- `formatOptions` — A string of JSON name-value pairs that provide additional options for formatting data at the destination. See Data format options.

Returns the `DataSink`.

def getSource

```
def getSource( connectionType : String,
               connectionOptions : JsonOptions,
               transformationContext : String = ""
               pushDownPredicate
             ) : DataSource
```

Creates a DataSource trait that reads data from a source like Amazon S3, JDBC, or the AWS Glue Data Catalog. Also supports Kafka and Kinesis streaming data sources.

- `connectionType` — The type of the data source. See the section called “Connection parameters”.

- `connectionOptions` — A string of JSON name-value pairs that provide additional information for establishing a connection with the data source. For more information, see the section called "Connection parameters".

  A Kinesis streaming source requires the following connection options: `streamARN`, `startingPosition`, `inferSchema`, and `classification`.

  A Kafka streaming source requires the following connection options: `connectionName`, `topicName`, `startingOffsets`, `inferSchema`, and `classification`.
- `transformationContext` — The transformation context that is associated with the sink to be used by job bookmarks. Set to empty by default.
- `pushDownPredicate` — Predicate on partition columns.

Returns the `DataSource`.

Example for Amazon Kinesis streaming source:

```
val kinesisOptions = jsonOptions()
data_frame_datasource0 = glueContext.getSource("kinesis",
 kinesisOptions).getDataFrame()

private def jsonOptions(): JsonOptions = {
    new JsonOptions(
      s"""{"streamARN": "arn:aws:kinesis:eu-central-1:123456789012:stream/
fromOptionsStream",
         |"startingPosition": "TRIM_HORIZON",
         |"inferSchema": "true",
         |"classification": "json"}""".stripMargin)
}
```

Example for Kafka streaming source:

```
val kafkaOptions = jsonOptions()
val data_frame_datasource0 = glueContext.getSource("kafka",
 kafkaOptions).getDataFrame()

private def jsonOptions(): JsonOptions = {
    new JsonOptions(
      s"""{"connectionName": "ConfluentKafka",
         |"topicName": "kafka-auth-topic",
        |"startingOffsets": "earliest",
        |"inferSchema": "true",
        |"classification": "json"}""".stripMargin)
}
```

def getSourceWithFormat

```
def getSourceWithFormat( connectionType : String,
                         options : JsonOptions,
                         transformationContext : String = "",
                         format : String = null,
                         formatOptions : JsonOptions = JsonOptions.empty
                       ) : DataSource
```

Creates a DataSource trait that reads data from a source like Amazon S3, JDBC, or the AWS Glue Data Catalog, and also sets the format of data stored in the source.

- `connectionType` – The type of the data source. See the section called “Connection parameters”.
- `options` – A string of JSON name-value pairs that provide additional information for establishing a connection with the data source. See the section called “Connection parameters”.
- `transformationContext` – The transformation context that is associated with the sink to be used by job bookmarks. Set to empty by default.
- `format` – The format of the data that is stored at the source. When the `connectionType` is "s3", you can also specify `format`. Can be one of “avro”, “csv”, “grokLog”, “ion”, “json”, “xml”, “parquet”, or “orc”.
- `formatOptions` – A string of JSON name-value pairs that provide additional options for parsing data at the source. See Data format options.

Returns the `DataSource`.

Examples

Create a DynamicFrame from a data source that is a comma-separated values (CSV) file on Amazon S3:

```
val datasource0 = glueContext.getSourceWithFormat(
    connectionType="s3",
    options =JsonOptions(s"""{"paths": [ "s3://csv/nycflights.csv"]}"""),
    transformationContext = "datasource0",
    format = "csv",
    formatOptions=JsonOptions(s"""{"withHeader":"true","separator": ","}""")
    ).getDynamicFrame()
```

Create a DynamicFrame from a data source that is a PostgreSQL using a JDBC connection:

```
val datasource0 = glueContext.getSourceWithFormat(
    connectionType="postgresql",
    options =JsonOptions(s"""{
      "url":"jdbc:postgresql://databasePostgres-1.rds.amazonaws.com:5432/testdb",
      "dbtable": "public.company",
      "redshiftTmpDir":"",
      "user":"username",
      "password":"password123"
    }"""),
    transformationContext = "datasource0").getDynamicFrame()
```

Create a DynamicFrame from a data source that is a MySQL using a JDBC connection:

```
 val datasource0 = glueContext.getSourceWithFormat(
    connectionType="mysql",
    options =JsonOptions(s"""{
      "url":"jdbc:mysql://databaseMysql-1.rds.amazonaws.com:3306/testdb",
      "dbtable": "athenatest_nycflights13_csv",
      "redshiftTmpDir":"",
      "user":"username",
      "password":"password123"
    }"""),
    transformationContext = "datasource0").getDynamicFrame()
```

def getSparkSession

```
def getSparkSession : SparkSession
```

Gets the `SparkSession` object associated with this GlueContext. Use this SparkSession object to register tables and UDFs for use with `DataFrame` created from DynamicFrames.

Returns the SparkSession.

def startTransaction

```
def startTransaction(readOnly: Boolean):String
```

Start a new transaction. Internally calls the Lake Formation startTransaction API.

- `readOnly` – (Boolean) Indicates whether this transaction should be read only or read and write. Writes made using a read-only transaction ID will be rejected. Read-only transactions do not need to be committed.

Returns the transaction ID.

def commitTransaction

```
def commitTransaction(transactionId: String, waitForCommit: Boolean): Boolean
```

Attempts to commit the specified transaction. `commitTransaction` may return before the transaction has finished committing. Internally calls the Lake Formation commitTransaction API.

- `transactionId` – (String) The transaction to commit.
- `waitForCommit` – (Boolean) Determines whether the `commitTransaction` returns immediately. The default value is true. If false, `commitTransaction` polls and waits until the transaction is committed. The amount of wait time is restricted to 1 minute using exponential backoff with a maximum of 6 retry attempts.

Returns a Boolean to indicate whether the commit is done or not.

def cancelTransaction

```
def cancelTransaction(transactionId: String): Unit
```

Attempts to cancel the specified transaction. Internally calls the Lake Formation CancelTransaction API.

- `transactionId` – (String) The transaction to cancel.

Returns a `TransactionCommittedException` exception if the transaction was previously committed.

def this

```
def this( sc : SparkContext,
          minPartitions : Int,
          targetPartitions : Int )
```

Creates a `GlueContext` object using the specified `SparkContext`, minimum partitions, and target partitions.

- `sc` — The `SparkContext`.
- `minPartitions` — The minimum number of partitions.
- `targetPartitions` — The target number of partitions.

Returns the `GlueContext`.

def this

```
def this( sc : SparkContext )
```

Creates a `GlueContext` object with the provided `SparkContext`. Sets the minimum partitions to 10 and target partitions to 20.

- `sc` — The `SparkContext`.

Returns the `GlueContext`.

def this

```
def this( sparkContext : JavaSparkContext )
```

Creates a `GlueContext` object with the provided `JavaSparkContext`. Sets the minimum partitions to 10 and target partitions to 20.

- `sparkContext` — The `JavaSparkContext`.

Returns the `GlueContext`.

MappingSpec

Package: com.amazonaws.services.glue

MappingSpec case class

```
case class MappingSpec( sourcePath: SchemaPath,
                        sourceType: DataType,
                        targetPath: SchemaPath,
                        targetType: DataTyp
                       ) extends Product4[String, String, String, String] {
  override def _1: String = sourcePath.toString
  override def _2: String = ExtendedTypeName.fromDataType(sourceType)
  override def _3: String = targetPath.toString
  override def _4: String = ExtendedTypeName.fromDataType(targetType)
}
```

- `sourcePath` — The `SchemaPath` of the source field.
- `sourceType` — The `DataType` of the source field.
- `targetPath` — The `SchemaPath` of the target field.
- `targetType` — The `DataType` of the target field.

A `MappingSpec` specifies a mapping from a source path and a source data type to a target path and a target data type. The value at the source path in the source frame appears in the target frame at the target path. The source data type is cast to the target data type.

It extends from `Product4` so that you can handle any `Product4` in your `applyMapping` interface.

MappingSpec object

```
object MappingSpec
```

The `MappingSpec` object has the following members:

Val orderingByTarget

```
val orderingByTarget: Ordering[MappingSpec]
```

Def apply

```
def apply( sourcePath : String,
           sourceType : DataType,
           targetPath : String,
           targetType : DataType
         ) : MappingSpec
```

Creates a `MappingSpec`.

- `sourcePath` — A string representation of the source path.
- `sourceType` — The source `DataType`.
- `targetPath` — A string representation of the target path.
- `targetType` — The target `DataType`.

Returns a `MappingSpec`.

Def apply

```
def apply( sourcePath : String,
           sourceTypeString : String,
           targetPath : String,
           targetTypeString : String
         ) : MappingSpec
```

Creates a `MappingSpec`.

- `sourcePath` — A string representation of the source path.
- `sourceType` — A string representation of the source data type.
- `targetPath` — A string representation of the target path.
- `targetType` — A string representation of the target data type.

Returns a MappingSpec.

Def apply

```
def apply( product : Product4[String, String, String, String] ) : MappingSpec
```

Creates a `MappingSpec`.

- `product` — The `Product4` of the source path, source data type, target path, and target data type.

Returns a `MappingSpec`.

AWS Glue Scala ResolveSpec APIs

Topics

- ResolveSpec object
- ResolveSpec case class

Package: com.amazonaws.services.glue

ResolveSpec object

ResolveSpec

```
object ResolveSpec
```

Def

```
def apply( path : String,
           action : String
         ) : ResolveSpec
```

Creates a `ResolveSpec`.

- `path` — A string representation of the choice field that needs to be resolved.
- `action` — A resolution action. The action can be one of the following: `Project`, `KeepAsStruct`, or `Cast`.

Returns the `ResolveSpec`.

Def

```
def apply( product : Product2[String, String] ) : ResolveSpec
```

Creates a `ResolveSpec`.

- `product` — `Product2` of: source path, resolution action.

Returns the `ResolveSpec`.

ResolveSpec case class

```
case class ResolveSpec extends Product2[String, String]  (
           path : SchemaPath,
           action : String )
```

Creates a `ResolveSpec`.

- `path` — The `SchemaPath` of the choice field that needs to be resolved.
- `action` — A resolution action. The action can be one of the following: `Project`, `KeepAsStruct`, or `Cast`.

ResolveSpec def methods

```
def _1 : String
```

```
def _2 : String
```

AWS Glue Scala ArrayNode APIs

Package: com.amazonaws.services.glue.types

ArrayNode case class

ArrayNode

```
case class ArrayNode extends DynamicNode  (
           value : ArrayBuffer[DynamicNode] )
```

ArrayNode def methods

```
def add( node : DynamicNode )
def clone
```

```
def equals( other : Any )
```

```
def get( index : Int ) : Option[DynamicNode]
```

```
def getValue
```

```
def hashCode : Int
```

```
def isEmpty : Boolean
```

```
def nodeType
```

```
def remove( index : Int )
```

```
def this
```

```
def toIterator : Iterator[DynamicNode]
```

```
def toJson : String
```

```
def update( index : Int,
            node : DynamicNode )
```

AWS Glue Scala BinaryNode APIs

Package: com.amazonaws.services.glue.types

BinaryNode case class

BinaryNode

```
case class BinaryNode extends ScalarNode(value, TypeCode.BINARY)  (
           value : Array[Byte] )
```

BinaryNode val fields

- `ordering`

BinaryNode def methods

```
def clone
```

```
def equals( other : Any )
```

```
def hashCode : Int
```

AWS Glue Scala BooleanNode APIs

Package: com.amazonaws.services.glue.types

BooleanNode case class

BooleanNode

```
case class BooleanNode extends ScalarNode(value, TypeCode.BOOLEAN)  (
           value : Boolean )
```

BooleanNode val fields

- `ordering`

BooleanNode def methods

```
def equals( other : Any )
```

AWS Glue Scala ByteNode APIs

Package: com.amazonaws.services.glue.types

ByteNode case class

ByteNode

```
case class ByteNode extends ScalarNode(value, TypeCode.BYTE)  (
           value : Byte )
```

ByteNode val fields

- `ordering`

ByteNode def methods

```
def equals( other : Any )
```

AWS Glue Scala DateNode APIs

Package: com.amazonaws.services.glue.types

DateNode case class

DateNode

```
case class DateNode extends ScalarNode(value, TypeCode.DATE)  (
           value : Date )
```

DateNode val fields

- `ordering`

DateNode def methods

```
def equals( other : Any )
```

```
def this( value : Int )
```

AWS Glue Scala DecimalNode APIs

Package: com.amazonaws.services.glue.types

DecimalNode case class

DecimalNode

```
case class DecimalNode extends ScalarNode(value, TypeCode.DECIMAL)  (
           value : BigDecimal )
```

DecimalNode val fields

- `ordering`

DecimalNode def methods

```
def equals( other : Any )
```

```
def this( value : Decimal )
```

AWS Glue Scala DoubleNode APIs

Package: com.amazonaws.services.glue.types

DoubleNode case class

DoubleNode

```
case class DoubleNode extends ScalarNode(value, TypeCode.DOUBLE)  (
           value : Double )
```

DoubleNode val fields

- `ordering`

DoubleNode def methods

```
def equals( other : Any )
```

AWS Glue Scala DynamicNode APIs

Topics

- DynamicNode class
- DynamicNode object

Package: com.amazonaws.services.glue.types

DynamicNode class

DynamicNode

```
class DynamicNode extends Serializable with Cloneable
```

DynamicNode def methods

```
def getValue : Any
```

Get plain value and bind to the current record:

```
def nodeType : TypeCode
```

```
def toJson : String
```

Method for debug:

```
def toRow( schema : Schema,
           options : Map[String, ResolveOption]
         ) : Row
```

```
def typeName : String
```

DynamicNode object

DynamicNode

```
object DynamicNode
```

DynamicNode def methods

```
def quote( field : String,
           useQuotes : Boolean
         ) : String
```

```
def quote( node : DynamicNode,
            useQuotes : Boolean
          ) : String
```

EvaluateDataQuality class

AWS Glue Data Quality is in preview release for AWS Glue and is subject to change.

Package: com.amazonaws.services.glue.dq

```
object EvaluateDataQuality
```

Def apply

```
def apply(frame: DynamicFrame,
            ruleset: String,
            publishingOptions: JsonOptions = JsonOptions.empty): DynamicFrame
```

Evaluates a data quality ruleset against a `DynamicFrame`, and returns a new `DynamicFrame` with results of the evaluation. To learn more about AWS Glue Data Quality, see AWS Glue Data Quality.

- `frame` – The `DynamicFrame` that you want to evaluate the data quality of.
- `ruleset` – A Data Quality Definition Language (DQDL) ruleset in string format. To learn more about DQDL, see the Data Quality Definition Language (DQDL) reference guide.
- `publishingOptions` – A dictionary that specifies the following options for publishing evaluation results and metrics:
  - `dataQualityEvaluationContext` – A string that specifies the namespace under which AWS Glue should publish Amazon CloudWatch metrics and the data quality results. The aggregated metrics appear in CloudWatch, while the full results appear in the AWS Glue Studio interface.
    - Required: No
    - Default value: `default_context`
  - `enableDataQualityCloudWatchMetrics` – Specifies whether the results of the data quality evaluation should be published to CloudWatch. You specify a namespace for the metrics using the `dataQualityEvaluationContext` option.
    - Required: No
    - Default value: False

- `enableDataQualityResultsPublishing` – Specifies whether the data quality results should be visible on the **Data Quality** tab in the AWS Glue Studio interface.
    - Required: No
    - Default value: True
- `resultsS3Prefix` – Specifies the Amazon S3 location where AWS Glue can write the data quality evaluation results.
    - Required: No
    - Default value: "" (empty string)

Example

The following example code demonstrates how to evaluate data quality for a `DynamicFrame` before performing a `SelectFields` transform. The script verifies that all data quality rules pass before it attempts the transform.

```
import com.amazonaws.services.glue.GlueContext
import com.amazonaws.services.glue.MappingSpec
import com.amazonaws.services.glue.errors.CallSite
import com.amazonaws.services.glue.util.GlueArgParser
import com.amazonaws.services.glue.util.Job
import com.amazonaws.services.glue.util.JsonOptions
import org.apache.spark.SparkContext
import scala.collection.JavaConverters._
import com.amazonaws.services.glue.dq.EvaluateDataQuality

object GlueApp {
  def main(sysArgs: Array[String]) {
    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)
    // @params: [JOB_NAME]
    val args = GlueArgParser.getResolvedOptions(sysArgs, Seq("JOB_NAME").toArray)
    Job.init(args("JOB_NAME"), glueContext, args.asJava)

    // Create DynamicFrame with data
    val Legislators_Area = glueContext.getCatalogSource(database="legislators",
 tableName="areas_json", transformationContext="S3bucket_node1").getDynamicFrame()

    // Define data quality ruleset
    val DQ_Ruleset = """
      Rules = [ColumnExists "id"]
    """

    // Evaluate data quality
    val DQ_Results = EvaluateDataQuality.apply(frame=Legislators_Area,
 ruleset=DQ_Ruleset, publishingOptions=JsonOptions("""{"dataQualityEvaluationContext":
 "Legislators_Area", "enableDataQualityMetrics": "true",
 "enableDataQualityResultsPublishing": "true"}"""))
    assert(DQ_Results.filter(_.getField("Outcome").contains("Failed")).count == 0,
 "Failing DQ rules for Legislators_Area caused the job to fail.")

    // Script generated for node Select Fields
    val SelectFields_Results = Legislators_Area.selectFields(paths=Seq("id", "name"),
 transformationContext="Legislators_Area")

    Job.commit()
  }
}
```

AWS Glue Scala FloatNode APIs

Package: com.amazonaws.services.glue.types

FloatNode case class

FloatNode

```
case class FloatNode extends ScalarNode(value, TypeCode.FLOAT)  (
           value : Float )
```

FloatNode val fields

- `ordering`

FloatNode def methods

```
def equals( other : Any )
```

FillMissingValues class

Package: com.amazonaws.services.glue.ml

```
object FillMissingValues
```

Def apply

```
def apply(frame: DynamicFrame,
          missingValuesColumn: String,
          outputColumn: String = "",
          transformationContext: String = "",
          callSite: CallSite = CallSite("Not provided", ""),
          stageThreshold: Long = 0,
          totalThreshold: Long = 0): DynamicFrame
```

Fills a dynamic frame's missing values in a specified column and returns a new frame with estimates in a new column. For rows without missing values, the specified column's value is duplicated to the new column.

- `frame` — The DynamicFrame in which to fill missing values. Required.
- `missingValuesColumn` — The column containing missing values (`null` values and empty strings). Required.
- `outputColumn` — The name of the new column that will contain estimated values for all rows whose value was missing. Optional; the default is the value of `missingValuesColumn` suffixed by `"_filled"`.
- `transformationContext` — A unique string that is used to identify state information (optional).
- `callSite` — Used to provide context information for error reporting. (optional).
- `stageThreshold` — The maximum number of errors that can occur in the transformation before it errors out (optional; the default is zero).
- `totalThreshold` — The maximum number of errors that can occur overall before processing errors out (optional; the default is zero).

Returns a new dynamic frame with one additional column that contains estimations for rows with missing values and the present value for other rows.

FindMatches class

Package: com.amazonaws.services.glue.ml

```
object FindMatches
```

Def apply

```
def apply(frame: DynamicFrame,
            transformId: String,
            transformationContext: String = "",
            callSite: CallSite = CallSite("Not provided", ""),
            stageThreshold: Long = 0,
            totalThreshold: Long = 0,
            enforcedMatches: DynamicFrame = null): DynamicFrame,
   computeMatchConfidenceScores: Boolean
```

Find matches in an input frame and return a new frame with a new column containing a unique ID per match group.

- `frame` — The DynamicFrame in which to find matches. Required.
- `transformId` — A unique ID associated with the FindMatches transform to apply on the input frame. Required.
- `transformationContext` — Identifier for this `DynamicFrame`. The `transformationContext` is used as a key for the job bookmark state that is persisted across runs. Optional.
- `callSite` — Used to provide context information for error reporting. These values are automatically set when calling from Python. Optional.
- `stageThreshold` — The maximum number of error records allowed from the computation of this `DynamicFrame` before throwing an exception, excluding records present in the previous `DynamicFrame`. Optional. The default is zero.
- `totalThreshold` — The maximum number of total errors records before an exception is thrown, including those from previous frames. Optional. The default is zero.
- `enforcedMatches` — The frame for enforced matches. Optional. The default is `null`.
- `computeMatchConfidenceScores` — A Boolean value indicating whether to compute a confidence score for each group of matching records. Optional. The default is false.

Returns a new dynamic frame with a unique identifier assigned to each group of matching records.

FindIncrementalMatches class

Package: com.amazonaws.services.glue.ml

```
object FindIncrementalMatches
```

Def apply

```
apply(existingFrame: DynamicFrame,
            incrementalFrame: DynamicFrame,
            transformId: String,
            transformationContext: String = "",
            callSite: CallSite = CallSite("Not provided", ""),
            stageThreshold: Long = 0,
            totalThreshold: Long = 0,
            enforcedMatches: DynamicFrame = null): DynamicFrame,
   computeMatchConfidenceScores: Boolean
```

Find matches across the existing and incremental frames and return a new frame with a column containing a unique ID per match group.

- `existingframe` — An existing frame which has been assigned a matching ID for each group. Required.
- `incrementalframe` — An incremental frame used to find matches against the existing frame. Required.
- `transformId` — A unique ID associated with the FindIncrementalMatches transform to apply on the input frames. Required.
- `transformationContext` — Identifier for this `DynamicFrame`. The `transformationContext` is used as a key for the job bookmark state that is persisted across runs. Optional.
- `callSite` — Used to provide context information for error reporting. These values are automatically set when calling from Python. Optional.
- `stageThreshold` — The maximum number of error records allowed from the computation of this `DynamicFrame` before throwing an exception, excluding records present in the previous `DynamicFrame`. Optional. The default is zero.
- `totalThreshold` — The maximum number of total errors records before an exception is thrown, including those from previous frames. Optional. The default is zero.
- `enforcedMatches` — The frame for enforced matches. Optional. The default is `null`.
- `computeMatchConfidenceScores` — A Boolean value indicating whether to compute a confidence score for each group of matching records. Optional. The default is false.

Returns a new dynamic frame with a unique identifier assigned to each group of matching records.

AWS Glue Scala IntegerNode APIs

Package: com.amazonaws.services.glue.types

IntegerNode case class

IntegerNode

```
case class IntegerNode extends ScalarNode(value, TypeCode.INT)  (
           value : Int )
```

IntegerNode val fields

- `ordering`

IntegerNode def methods

```
def equals( other : Any )
```

AWS Glue Scala LongNode APIs

Package: com.amazonaws.services.glue.types

LongNode case class

LongNode

```
case class LongNode extends ScalarNode(value, TypeCode.LONG)  (
           value : Long )
```

LongNode val fields

- `ordering`

LongNode def methods

```
def equals( other : Any )
```

AWS Glue Scala MapLikeNode APIs

Package: com.amazonaws.services.glue.types

MapLikeNode class

MapLikeNode

```
class MapLikeNode extends DynamicNode  (
           value : mutable.Map[String, DynamicNode] )
```

MapLikeNode def methods

```
def clear : Unit
```

```
def get( name : String ) : Option[DynamicNode]
```

```
def getValue
```

```
def has( name : String ) : Boolean
```

```
def isEmpty : Boolean
```

```
def put( name : String,
         node : DynamicNode
       ) : Option[DynamicNode]
```

```
def remove( name : String ) : Option[DynamicNode]
```

```
def toIterator : Iterator[(String, DynamicNode)]
```

```
def toJson : String
```

```
def toJson( useQuotes : Boolean ) : String
```

Example: Given this JSON:

```
{"foo": "bar"}
```

If `useQuotes == true`, `toJson` yields `{"foo": "bar"}`. If `useQuotes == false`, `toJson` yields `{foo: bar}` @return.

AWS Glue Scala MapNode APIs

Package: com.amazonaws.services.glue.types

MapNode case class

MapNode

```
case class MapNode extends MapLikeNode(value)  (
           value : mutable.Map[String, DynamicNode] )
```

MapNode def methods

```
def clone
```

```
def equals( other : Any )
```

```
def hashCode : Int
```

```
def nodeType
```

```
def this
```

AWS Glue Scala NullNode APIs

Topics

- NullNode class
- NullNode case object

Package: com.amazonaws.services.glue.types

NullNode class

NullNode

```
class NullNode
```

NullNode case object

NullNode

```
case object NullNode extends NullNode
```

AWS Glue Scala ObjectNode APIs

Topics

- ObjectNode object
- ObjectNode case class

Package: com.amazonaws.services.glue.types

ObjectNode object

ObjectNode

```
object ObjectNode
```

ObjectNode def methods

```
def apply( frameKeys : Set[String],
           v1 : mutable.Map[String, DynamicNode],
           v2 : mutable.Map[String, DynamicNode],
           resolveWith : String
         ) : ObjectNode
```

ObjectNode case class

ObjectNode

```
case class ObjectNode extends MapLikeNode(value)  (
            val value : mutable.Map[String, DynamicNode] )
```

ObjectNode def methods

```
def clone
```

```
def equals( other : Any )
```

```
def hashCode : Int
```

```
def nodeType
```

```
def this
```

AWS Glue Scala ScalarNode APIs

Topics

- ScalarNode class
- ScalarNode object

Package: com.amazonaws.services.glue.types

ScalarNode class

ScalarNode

```
class ScalarNode extends DynamicNode  (
           value : Any,
           scalarType : TypeCode )
```

ScalarNode def methods

```
def compare( other : Any,
             operator : String
           ) : Boolean
```

```
def getValue
def hashCode : Int
```

```
def nodeType
```

```
def toJson
```

ScalarNode object

ScalarNode

```
object ScalarNode
```

ScalarNode def methods

```
def apply( v : Any ) : DynamicNode
```

```
def compare( tv : Ordered[T],
             other : T,
             operator : String
           ) : Boolean
```

```
def compareAny( v : Any,
                y : Any,
                o : String )
```

```
def withEscapedSpecialCharacters( jsonToEscape : String ) : String
```

AWS Glue Scala ShortNode APIs

Package: com.amazonaws.services.glue.types

ShortNode case class

ShortNode

```
case class ShortNode extends ScalarNode(value, TypeCode.SHORT)  (
           value : Short )
```

ShortNode val fields

- `ordering`

ShortNode def methods

```
def equals( other : Any )
```

AWS Glue Scala StringNode APIs

Package: com.amazonaws.services.glue.types

StringNode case class

StringNode

```
case class StringNode extends ScalarNode(value, TypeCode.STRING)  (
           value : String )
```

StringNode val fields

- `ordering`

StringNode def methods

```
def equals( other : Any )
```

```
def this( value : UTF8String )
```

AWS Glue Scala TimestampNode APIs

Package: com.amazonaws.services.glue.types

TimestampNode case class

TimestampNode

```
case class TimestampNode extends ScalarNode(value, TypeCode.TIMESTAMP)  (
           value : Timestamp )
```

TimestampNode val fields

- `ordering`

TimestampNode def methods

```
def equals( other : Any )
```

```
def this( value : Long )
```

AWS Glue Scala GlueArgParser APIs

Package: com.amazonaws.services.glue.util

GlueArgParser object

GlueArgParser

```
object GlueArgParser
```

This is strictly consistent with the Python version of `utils.getResolvedOptions` in the `AWSGlueDataplanePython` package.

GlueArgParser def methods

```
def getResolvedOptions( args : Array[String],
                        options : Array[String]
                      ) : Map[String, String]
```

```
def initParser( userOptionsSet : mutable.Set[String] ) : ArgumentParser
```

Example Retrieving arguments passed to a job

To retrieve job arguments, you can use the `getResolvedOptions` method. Consider the following example, which retrieves a job argument named `aws_region`.

```
val args = GlueArgParser.getResolvedOptions(sysArgs,
 Seq("JOB_NAME","aws_region").toArray)
Job.init(args("JOB_NAME"), glueContext, args.asJava)
val region = args("aws_region")
```

```
println(region)
```

AWS Glue Scala job APIs

Package: com.amazonaws.services.glue.util

Job object

Job

```
object Job
```

Job def methods

```
def commit
```

```
def init( jobName : String,
          glueContext : GlueContext,
          args : java.util.Map[String, String] = Map[String, String]().asJava
        ) : this.type
```

```
def init( jobName : String,
          glueContext : GlueContext,
          endpoint : String,
          args : java.util.Map[String, String]
        ) : this.type
```

```
def isInitialized
```

```
def reset
```

```
def runId
```

Features and optimizations for programming AWS Glue for Spark ETL scripts

The following sections describe techniques and values that apply generally to AWS Glue for Spark ETL (extract, transform, and load) programming in any language.

Topics

- Connection types and options for ETL in AWS Glue for Spark
- Data format options for inputs and outputs in AWS Glue for Spark
- AWS Glue Data Catalog support for Spark SQL jobs
- Using job bookmarks
- Using Sensitive Data Detection outside AWS Glue Studio
- AWS Glue Visual Job API

Connection types and options for ETL in AWS Glue for Spark

In AWS Glue for Spark, various PySpark and Scala methods and transforms specify the connection type using a `connectionType` parameter. They specify connection options using a `connectionOptions` or `options` parameter.

The `connectionType` parameter can take the values shown in the following table. The associated `connectionOptions` (or `options`) parameter values for each type are documented in the following sections. Except where otherwise noted, the parameters apply when the connection is used as a source or sink.

For sample code that demonstrates setting and using connection options, see the homepage for each connection type.

`connectionType`	Connects to
dynamodb	Amazon DynamoDB database
kinesis	Amazon Kinesis Data Streams
s3	Amazon S3
documentdb	Amazon DocumentDB (with MongoDB compatibility) database
opensearch	Amazon OpenSearch Service.
redshift	Amazon Redshift database
kafka	Kafka or Amazon Managed Streaming for Apache Kafka

`connectionType`	Connects to
azurecosmos	Azure Cosmos for NoSQL.
azuresql	Azure SQL.
bigquery	Google BigQuery.
mongodb	MongoDB database, including MongoDB Atlas.
sqlserver	Microsoft SQL Server database (see JDBC connections)
mysql	MySQL database (see JDBC connections)
oracle	Oracle database (see JDBC connections)
postgresql	PostgreSQL database (see JDBC connections)
saphana	SAP HANA.
snowflake	Snowflake data lake
teradata	Teradata Vantage.
vertica	Vertica.
custom.*	Spark, Athena, or JDBC data stores (see Custom and AWS Marketplace connectionType values
marketplace.*	Spark, Athena, or JDBC data stores (see Custom and AWS Marketplace connectionType values)

DynamoDB connections

You can use AWS Glue for Spark to read from and write to tables in DynamoDB in AWS Glue. You connect to DynamoDB using IAM permissions attached to your AWS Glue job. AWS Glue supports writing data into another AWS account's DynamoDB table. For more information, see the section called “Cross-account cross-Region access to DynamoDB tables”.

In addition to the AWS Glue DynamoDB ETL connector, you can read from DynamoDB using the DynamoDB export connector, that invokes a DynamoDB `ExportTableToPointInTime` request and stores it in an Amazon S3 location you supply, in the format of DynamoDB JSON. AWS Glue then creates a DynamicFrame object by reading the data from the Amazon S3 export location.

The DynamoDB writer is available in AWS Glue version 1.0 or later versions. The AWS Glue DynamoDB export connector is available in AWS Glue version 2.0 or later versions.

For more information about DynamoDB, consult Amazon DynamoDB documentation.

> **Note**
>
> The DynamoDB ETL reader does not support filters or pushdown predicates.

Configuring DynamoDB connections

To connect to DynamoDB from AWS Glue, grant the IAM role associated with your AWS Glue job permission to interact with DynamoDB. For more information about permissions necessary to read or write from DynamoDB, consult Actions, resources, and condition keys for DynamoDB in the IAM documentation.

In the following situations, you may need additional configuration:

- When using the DynamoDB export connector, you will need to configure IAM so your job can request DynamoDB table exports. Additionally, you will need to identify an Amazon S3 bucket for the export and provide appropriate permissions in IAM for DynamoDB to write to it, and for your AWS Glue job to read from it. For more information, consult Request a table export in DynamoDB.
- If your AWS Glue job has specific Amazon VPC connectivity requirements, use the `NETWORK` AWS Glue connection type to provide network options. Since access to DynamoDB is authorized by IAM, there is no need for a AWS Glue DynamoDB connection type.

Reading from and writing to DynamoDB

The following code examples show how to read from (via the ETL connector) and write to DynamoDB tables. They demonstrate reading from one table and writing to another table.

Python

```
import sys
```

```
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job
from awsglue.utils import getResolvedOptions

args = getResolvedOptions(sys.argv, ["JOB_NAME"])
glue_context= GlueContext(SparkContext.getOrCreate())
job = Job(glue_context)
job.init(args["JOB_NAME"], args)

dyf = glue_context.create_dynamic_frame.from_options(
    connection_type="dynamodb",
    connection_options={"dynamodb.input.tableName": test_source,
        "dynamodb.throughput.read.percent": "1.0",
        "dynamodb.splits": "100"
    }
)
print(dyf.getNumPartitions())

glue_context.write_dynamic_frame_from_options(
    frame=dyf,
    connection_type="dynamodb",
    connection_options={"dynamodb.output.tableName": test_sink,
        "dynamodb.throughput.write.percent": "1.0"
    }
)

job.commit()
```

Scala

```
import com.amazonaws.services.glue.GlueContext
import com.amazonaws.services.glue.util.GlueArgParser
import com.amazonaws.services.glue.util.Job
import com.amazonaws.services.glue.util.JsonOptions
import com.amazonaws.services.glue.DynamoDbDataSink
import org.apache.spark.SparkContext
import scala.collection.JavaConverters._


object GlueApp {

  def main(sysArgs: Array[String]): Unit = {
    val glueContext = new GlueContext(SparkContext.getOrCreate())
    val args = GlueArgParser.getResolvedOptions(sysArgs, Seq("JOB_NAME").toArray)
    Job.init(args("JOB_NAME"), glueContext, args.asJava)

    val dynamicFrame = glueContext.getSourceWithFormat(
      connectionType = "dynamodb",
      options = JsonOptions(Map(
        "dynamodb.input.tableName" -> test_source,
        "dynamodb.throughput.read.percent" -> "1.0",
        "dynamodb.splits" -> "100"
      ))
    ).getDynamicFrame()

    print(dynamicFrame.getNumPartitions())

    val dynamoDbSink: DynamoDbDataSink =  glueContext.getSinkWithFormat(
      connectionType = "dynamodb",
      options = JsonOptions(Map(
        "dynamodb.output.tableName" -> test_sink,
        "dynamodb.throughput.write.percent" -> "1.0"
      ))
    ).asInstanceOf[DynamoDbDataSink]

    dynamoDbSink.writeDynamicFrame(dynamicFrame)

    Job.commit()
  }

}
```

Using the DynamoDB export connector

The export connector performs better than the ETL connector when the DynamoDB table size is larger than 80 GB. In addition, given that the export request is conducted outside from the Spark processes in an AWS Glue job, you can enable auto scaling of AWS Glue jobs to save DPU usage during the export request. With the export connector, you also do not need to configure the number of splits for Spark executor parallelism or DynamoDB throughput read percentage.

Note

DynamoDB has specific requirements to invoke the `ExportTableToPointInTime` requests. For more information, see Requesting a table export in DynamoDB. For example, Point-in-Time-Restore (PITR) needs to be enabled on the table to use this connector.
The DynamoDB connector also supports AWS KMS encryption for DynamoDB exports to Amazon S3. Supplying your security configuration in the AWS Glue job configuration enables AWS KMS encryption for a DynamoDB export. The KMS key must be in the same Region as the Amazon S3 bucket.
Note that additional charges for DynamoDB export and Amazon S3 storage costs apply. Exported data in Amazon S3 persists after a job run finishes so you can reuse it without additional DynamoDB exports. A requirement for using this connector is that point-in-time recovery (PITR) is enabled for the table.
The DynamoDB ETL connector or export connector do not support filters or pushdown predicates to be applied at the DynamoDB source.

The following code examples show how to read from (via the export connector) and print the number of partitions.

Python

```
import sys
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job
from awsglue.utils import getResolvedOptions

args = getResolvedOptions(sys.argv, ["JOB_NAME"])
glue_context= GlueContext(SparkContext.getOrCreate())
job = Job(glue_context)
job.init(args["JOB_NAME"], args)

dyf = glue_context.create_dynamic_frame.from_options(
    connection_type="dynamodb",
    connection_options={
        "dynamodb.export": "ddb",
        "dynamodb.tableArn": test_source,
        "dynamodb.s3.bucket": bucket_name,
        "dynamodb.s3.prefix": bucket_prefix,
        "dynamodb.s3.bucketOwner": account_id_of_bucket,
    }
)
print(dyf.getNumPartitions())

job.commit()
```

Scala

```
import com.amazonaws.services.glue.GlueContext
import com.amazonaws.services.glue.util.GlueArgParser
import com.amazonaws.services.glue.util.Job
import com.amazonaws.services.glue.util.JsonOptions
import com.amazonaws.services.glue.DynamoDbDataSink
import org.apache.spark.SparkContext
import scala.collection.JavaConverters._


object GlueApp {

  def main(sysArgs: Array[String]): Unit = {
    val glueContext = new GlueContext(SparkContext.getOrCreate())
    val args = GlueArgParser.getResolvedOptions(sysArgs, Seq("JOB_NAME").toArray)
    Job.init(args("JOB_NAME"), glueContext, args.asJava)

    val dynamicFrame = glueContext.getSourceWithFormat(
      connectionType = "dynamodb",
      options = JsonOptions(Map(
        "dynamodb.export" -> "ddb",
        "dynamodb.tableArn" -> test_source,
        "dynamodb.s3.bucket" -> bucket_name,
        "dynamodb.s3.prefix" -> bucket_prefix,
        "dynamodb.s3.bucketOwner" -> account_id_of_bucket,
      ))
    ).getDynamicFrame()

    print(dynamicFrame.getNumPartitions())

    Job.commit()
  }

}
```

These examples show how to do the read from (via the export connector) and print the number of partitions from an AWS Glue Data Catalog table that has a `dynamodb` classification:

Python

```
import sys
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job
from awsglue.utils import getResolvedOptions

args = getResolvedOptions(sys.argv, ["JOB_NAME"])
glue_context= GlueContext(SparkContext.getOrCreate())
job = Job(glue_context)
job.init(args["JOB_NAME"], args)

dynamicFrame = glue_context.create_dynamic_frame.from_catalog(
        database=catalog_database,
        table_name=catalog_table_name,
        additional_options={
            "dynamodb.export": "ddb",
            "dynamodb.s3.bucket": s3_bucket,
            "dynamodb.s3.prefix": s3_bucket_prefix
        }
    )
print(dynamicFrame.getNumPartitions())

job.commit()
```

Scala

```
import com.amazonaws.services.glue.GlueContext
import com.amazonaws.services.glue.util.GlueArgParser
import com.amazonaws.services.glue.util.Job
import com.amazonaws.services.glue.util.JsonOptions
import com.amazonaws.services.glue.DynamoDbDataSink
import org.apache.spark.SparkContext
import scala.collection.JavaConverters._


object GlueApp {

  def main(sysArgs: Array[String]): Unit = {
    val glueContext = new GlueContext(SparkContext.getOrCreate())
    val args = GlueArgParser.getResolvedOptions(sysArgs, Seq("JOB_NAME").toArray)
    Job.init(args("JOB_NAME"), glueContext, args.asJava)

    val dynamicFrame = glueContext.getCatalogSource(
        database = catalog_database,
        tableName = catalog_table_name,
        additionalOptions = JsonOptions(Map(
            "dynamodb.export" -> "ddb",
            "dynamodb.s3.bucket" -> s3_bucket,
            "dynamodb.s3.prefix" -> s3_bucket_prefix
        ))
    ).getDynamicFrame()
    print(dynamicFrame.getNumPartitions())
)
```

Simplifying usage of DynamoDB export JSON

The DynamoDB exports with the AWS Glue DynamoDB export connector results in JSON files of specific nested structures. For more information, see Data objects. AWS Glue supplies a DynamicFrame transformation, which can unnest such structures into an easier-to-use form for downstream applications.

The transform can be invoked in one of two ways. You can set the connection option `"dynamodb.simplifyDDBJson"` with the value `"true"` when calling a method to read from DynamoDB. You can also call the transform as a method independently available in the AWS Glue library.

Consider the following schema generated by a DynamoDB export:

```
root
|-- Item: struct
|    |-- parentMap: struct
|    |    |-- M: struct
|    |    |    |-- childMap: struct
|    |    |    |    |-- M: struct
|    |    |    |    |    |-- appName: struct
|    |    |    |    |    |    |-- S: string
|    |    |    |    |    |-- packageName: struct
|    |    |    |    |    |    |-- S: string
|    |    |    |    |    |-- updatedAt: struct
|    |    |    |    |    |    |-- N: string
|    |-- strings: struct
|    |    |-- SS: array
|    |    |    |-- element: string
|    |-- numbers: struct
|    |    |-- NS: array
|    |    |    |-- element: string
|    |-- binaries: struct
|    |    |-- BS: array
|    |    |    |-- element: string
|    |-- isDDBJson: struct
|    |    |-- BOOL: boolean
|    |-- nullValue: struct
|    |    |-- NULL: boolean
```

The `simplifyDDBJson` transform will simplify this to:

```
root
|-- parentMap: struct
|    |-- childMap: struct
|    |    |-- appName: string
|    |    |-- packageName: string
|    |    |-- updatedAt: string
|-- strings: array
|    |-- element: string
|-- numbers: array
|    |-- element: string
|-- binaries: array
|    |-- element: string
|-- isDDBJson: boolean
|-- nullValue: null
```

Note

`simplifyDDBJson` is available in AWS Glue 3.0 and later versions. The `unnestDDBJson` transform is also available to simplify DynamoDB export JSON. We encourage users to transition to `simplifyDDBJson` from `unnestDDBJson`.

Configuring paralleism in DynamoDB operations

To improve performance, you can tune certain parameters available for the DynamoDB connector. Your goal when tuning paralleism parameters is to maximize the use of the provisioned AWS

Glue workers. Then, if you need more performance, we recommend you to scale out your job by increasing the number of DPUs.

You can alter the parallelism in a DynamoDB read operation using the `dynamodb.splits` parameter when using the ETL connector. When reading with the export connector, you do not need to configure the number of splits for Spark executor parallelism. You can alter the parallelism in a DynamoDB write operation with `dynamodb.output.numParallelTasks`.

Reading with the DynamoDB ETL connector

We recommend you to calculate `dynamodb.splits` based on the maximum number of workers set in your job configuration and the following `numSlots` calculation. If autoscaling, the actual number of workers available may change under that cap. For more information about setting the maximum number of workers, see **Number of workers** (`NumberOfWorkers`) in the section called "Configuring Spark job properties".

- `numExecutors = NumberOfWorkers - 1`

  For context, one executor is reserved for the Spark driver; other executors are used to process data.
- `numSlotsPerExecutor =`

  AWS Glue 3.0 and later versions
  - 4 if `WorkerType` is `G.1X`
  - 8 if `WorkerType` is `G.2X`
  - 16 if `WorkerType` is `G.4X`
  - 32 if `WorkerType` is `G.8X`

  AWS Glue 2.0 and legacy versions
  - 8 if `WorkerType` is `G.1X`
  - 16 if `WorkerType` is `G.2X`
- `numSlots = numSlotsPerExecutor * numExecutors`

We recommend you set `dynamodb.splits` to the number of slots available, `numSlots`.

Writing to DynamoDB

The `dynamodb.output.numParallelTasks` parameter is used to determine WCU per Spark task, using the following calculation:

`permittedWcuPerTask = ( TableWCU * dynamodb.throughput.write.percent ) / dynamodb.output.numParallelTasks`

The DynamoDB writer will function best if configuration accurately represents the number of Spark tasks writing to DynamoDB. In some cases, you may need to override the default calculation to improve write performance. If you do not specify this parameter, the permitted WCU per Spark task will be automatically calculated by the following formula:

-
  - `numPartitions = dynamicframe.getNumPartitions()`
  - `numSlots` (as defined previously in this section)
  - `numParallelTasks = min(numPartitions, numSlots)`
- Example 1. DPU=10, WorkerType=Standard. Input DynamicFrame has 100 RDD partitions.
  - `numPartitions = 100`
  - `numExecutors = (10 - 1) * 2 - 1 = 17`
  - `numSlots = 4 * 17 = 68`
  - `numParallelTasks = min(100, 68) = 68`
- Example 2. DPU=10, WorkerType=Standard. Input DynamicFrame has 20 RDD partitions.
  - `numPartitions = 20`
  - `numExecutors = (10 - 1) * 2 - 1 = 17`
  - `numSlots = 4 * 17 = 68`
  - `numParallelTasks = min(20, 68) = 20`

> **Note**
>
> Jobs on legacy AWS Glue versions and those using Standard workers require different methods to calculate the number of slots. If you need to performance tune these jobs, we recommend you transition to supported AWS Glue versions.

DynamoDB connection option reference

Designates a connection to Amazon DynamoDB.

Connection options differ for a source connection and a sink connection.

"connectionType": "dynamodb" with the ETL connector as source

Use the following connection options with `"connectionType": "dynamodb"` as a source, when using the AWS Glue DynamoDB ETL connector:

- `"dynamodb.input.tableName"`: (Required) The DynamoDB table to read from.
- `"dynamodb.throughput.read.percent"`: (Optional) The percentage of read capacity units (RCU) to use. The default is set to "0.5". Acceptable values are from "0.1" to "1.5", inclusive.
  - `0.5` represents the default read rate, meaning that AWS Glue will attempt to consume half of the read capacity of the table. If you increase the value above `0.5`, AWS Glue increases the request rate; decreasing the value below `0.5` decreases the read request rate. (The actual read rate will vary, depending on factors such as whether there is a uniform key distribution in the DynamoDB table.)
  - When the DynamoDB table is in on-demand mode, AWS Glue handles the read capacity of the table as 40000. For exporting a large table, we recommend switching your DynamoDB table to on-demand mode.
- `"dynamodb.splits"`: (Optional) Defines how many splits we partition this DynamoDB table into while reading. The default is set to "1". Acceptable values are from "1" to "1,000,000", inclusive.

  1 represents there is no parallelism. We highly recommend that you specify a larger value for better performance by using the below formula. For more information on appropriately setting a value, see the section called "DynamoDB parallelism".
- `"dynamodb.sts.roleArn"`: (Optional) The IAM role ARN to be assumed for cross-account access. This parameter is available in AWS Glue 1.0 or later.
- `"dynamodb.sts.roleSessionName"`: (Optional) STS session name. The default is set to "glue-dynamodb-read-sts-session". This parameter is available in AWS Glue 1.0 or later.

"connectionType": "dynamodb" with the AWS Glue DynamoDB export connector as source

Use the following connection options with "connectionType": "dynamodb" as a source, when using the AWS Glue DynamoDB export connector, which is available only for AWS Glue version 2.0 onwards:

- `"dynamodb.export"`: (Required) A string value:
  - If set to ddb enables the AWS Glue DynamoDB export connector where a new `ExportTableToPointInTimeRequest` will be invoked during the AWS Glue job. A

  new export will be generated with the location passed from `dynamodb.s3.bucket` and `dynamodb.s3.prefix`.

  - If set to `s3` enables the AWS Glue DynamoDB export connector but skips the creation of a new DynamoDB export and instead uses the `dynamodb.s3.bucket` and `dynamodb.s3.prefix` as the Amazon S3 location of a past export of that table.

- `"dynamodb.tableArn"`: (Required) The DynamoDB table to read from.
- `"dynamodb.unnestDDBJson"`: (Optional) Default: false. Valid values: boolean. If set to true, performs an unnest transformation of the DynamoDB JSON structure that is present in exports. It is an error to set `"dynamodb.unnestDDBJson"` and `"dynamodb.simplifyDDBJson"` to true at the same time. In AWS Glue 3.0 and later versions, we recommend you use `"dynamodb.simplifyDDBJson"` for better behavior when simplifying DynamoDB Map types. For more information, see the section called “Simplifying usage of DynamoDB export JSON”.
- `"dynamodb.simplifyDDBJson"`: (Optional) Default: false. Valid values: boolean. If set to true, performs a transformation to simplify the schema of the DynamoDB JSON structure that is present in exports. This has the same purpose as the `"dynamodb.unnestDDBJson"` option but provides better support for DynamoDB Map types or even nested Map types in your DynamoDB table. This option is available in AWS Glue 3.0 and later versions. It is an error to set `"dynamodb.unnestDDBJson"` and `"dynamodb.simplifyDDBJson"` to true at the same time. For more information, see the section called “Simplifying usage of DynamoDB export JSON”.
- `"dynamodb.s3.bucket"`: (Optional) Indicates the Amazon S3 bucket location in which the DynamoDB `ExportTableToPointInTime` process is to be conducted. The file format for the export is DynamoDB JSON.
  - `"dynamodb.s3.prefix"`: (Optional) Indicates the Amazon S3 prefix location inside the Amazon S3 bucket in which the DynamoDB `ExportTableToPointInTime` loads are to be stored. If neither `dynamodb.s3.prefix` nor `dynamodb.s3.bucket` are specified, these values will default to the Temporary Directory location specified in the AWS Glue job configuration. For more information, see Special Parameters Used by AWS Glue.
  - `"dynamodb.s3.bucketOwner"`: Indicates the bucket owner needed for cross-account Amazon S3 access.
- `"dynamodb.sts.roleArn"`: (Optional) The IAM role ARN to be assumed for cross-account access and/or cross-Region access for the DynamoDB table. Note: The same IAM role ARN will be used to access the Amazon S3 location specified for the `ExportTableToPointInTime` request.
- `"dynamodb.sts.roleSessionName"`: (Optional) STS session name. The default is set to "glue-dynamodb-read-sts-session".

- `"dynamodb.exportTime"` (Optional) Valid values: strings representing ISO-8601 instants. A point-in-time at which the export should be made.
- `"dynamodb.sts.region"`: (Required if making a cross-region call using a regional endpoint) The region hosting the DynamoDB table you want to read.

"connectionType": "dynamodb" with the ETL connector as sink

Use the following connection options with `"connectionType": "dynamodb"` as a sink:

- `"dynamodb.output.tableName"`: (Required) The DynamoDB table to write to.
- `"dynamodb.throughput.write.percent"`: (Optional) The percentage of write capacity units (WCU) to use. The default is set to "0.5". Acceptable values are from "0.1" to "1.5", inclusive.
  - `0.5` represents the default write rate, meaning that AWS Glue will attempt to consume half of the write capacity of the table. If you increase the value above 0.5, AWS Glue increases the request rate; decreasing the value below 0.5 decreases the write request rate. (The actual write rate will vary, depending on factors such as whether there is a uniform key distribution in the DynamoDB table).
  - When the DynamoDB table is in on-demand mode, AWS Glue handles the write capacity of the table as `40000`. For importing a large table, we recommend switching your DynamoDB table to on-demand mode.
- `"dynamodb.output.numParallelTasks"`: (Optional) Defines how many parallel tasks write into DynamoDB at the same time. Used to calculate permissive WCU per Spark task. In most cases, AWS Glue will calculate a reasonable default for this value. For more information, see the section called “DynamoDB parallelism”.
- `"dynamodb.output.retry"`: (Optional) Defines how many retries we perform when there is a `ProvisionedThroughputExceededException` from DynamoDB. The default is set to "10".
- `"dynamodb.sts.roleArn"`: (Optional) The IAM role ARN to be assumed for cross-account access.
- `"dynamodb.sts.roleSessionName"`: (Optional) STS session name. The default is set to "glue-dynamodb-write-sts-session".

Cross-account cross-Region access to DynamoDB tables

AWS Glue ETL jobs support both cross-region and cross-account access to DynamoDB tables. AWS Glue ETL jobs support both reading data from another AWS Account's DynamoDB table, and writing data into another AWS Account's DynamoDB table. AWS Glue also supports both reading from a DynamoDB table in another region, and writing into a DynamoDB table in another region. This section gives instructions on setting up the access, and provides an example script.

The procedures in this section reference an IAM tutorial for creating an IAM role and granting access to the role. The tutorial also discusses assuming a role, but here you will instead use a job script to assume the role in AWS Glue. This tutorial also contains information about general cross-account practices. For more information, see Tutorial: Delegate Access Across AWS Accounts Using IAM Roles in the *IAM User Guide*.

Create a role

Follow step 1 in the tutorial to create an IAM role in account A. When defining the permissions of the role, you can choose to attach existing policies such as `AmazonDynamoDBReadOnlyAccess`, or `AmazonDynamoDBFullAccess` to allow the role to read/write DynamoDB. The following example shows creating a role named `DynamoDBCrossAccessRole`, with the permission policy `AmazonDynamoDBFullAccess`.

Grant access to the role

Follow step 2 in the tutorial in the *IAM User Guide* to allow account B to switch to the newly-created role. The following example creates a new policy with the following statement:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": {
    "Effect": "Allow",
    "Action": "sts:AssumeRole",
    "Resource": "arn:aws:iam::111122223333:role/DynamoDBCrossAccessRole"
  }
}
```

Then, you can attach this policy to the group/role/user you would like to use to access DynamoDB.

Assume the role in the AWS Glue job script

Now, you can log in to account B and create an AWS Glue job. To create a job, refer to the instructions at Configuring job properties for Spark jobs in AWS Glue.

In the job script you need to use the `dynamodb.sts.roleArn` parameter to assume the `DynamoDBCrossAccessRole` role. Assuming this role allows you to get the temporary credentials, which need to be used to access DynamoDB in account B. Review these example scripts.

For a cross-account read across regions (ETL connector):

```
import sys
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job
from awsglue.utils import getResolvedOptions

args = getResolvedOptions(sys.argv, ["JOB_NAME"])
glue_context= GlueContext(SparkContext.getOrCreate())
job = Job(glue_context)
job.init(args["JOB_NAME"], args)

dyf = glue_context.create_dynamic_frame_from_options(
    connection_type="dynamodb",
    connection_options={
    "dynamodb.region": "us-east-1",
    "dynamodb.input.tableName": "test_source",
    "dynamodb.sts.roleArn": "<DynamoDBCrossAccessRole's ARN>"
    }
)
dyf.show()
job.commit()
```

For a cross-account read across regions (ELT connector):

```
import sys
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job
from awsglue.utils import getResolvedOptions

args = getResolvedOptions(sys.argv, ["JOB_NAME"])
glue_context= GlueContext(SparkContext.getOrCreate())
job = Job(glue_context)
job.init(args["JOB_NAME"], args)

dyf = glue_context.create_dynamic_frame_from_options(
    connection_type="dynamodb",
    connection_options={
        "dynamodb.export": "ddb",
        "dynamodb.tableArn": "<test_source ARN>",
        "dynamodb.sts.roleArn": "<DynamoDBCrossAccessRole's ARN>"
    }
)
dyf.show()
job.commit()
```

For a read and cross-account write across regions:

```
import sys
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job
from awsglue.utils import getResolvedOptions

args = getResolvedOptions(sys.argv, ["JOB_NAME"])
glue_context= GlueContext(SparkContext.getOrCreate())
job = Job(glue_context)
job.init(args["JOB_NAME"], args)

dyf = glue_context.create_dynamic_frame_from_options(
    connection_type="dynamodb",
    connection_options={
        "dynamodb.region": "us-east-1",
        "dynamodb.input.tableName": "test_source"
    }
)
dyf.show()

glue_context.write_dynamic_frame_from_options(
    frame=dyf,
    connection_type="dynamodb",
    connection_options={
        "dynamodb.region": "us-west-2",
        "dynamodb.output.tableName": "test_sink",
        "dynamodb.sts.roleArn": "<DynamoDBCrossAccessRole's ARN>"
    }
)

job.commit()
```

Kinesis connections

You can use a Kinesis connection to read and write to Amazon Kinesis data streams using information stored in a Data Catalog table, or by providing information to directly access the data stream. You can read information from Kinesis into a Spark DataFrame, then convert it to a AWS Glue DynamicFrame. You can write DynamicFrames to Kinesis in a JSON format. If you directly access the data stream, use these options to provide the information about how to access the data stream.

If you use `getCatalogSource` or `create_data_frame_from_catalog` to consume records from a Kinesis streaming source, the job has the Data Catalog database and table name information, and can use that to obtain some basic parameters for reading from the Kinesis streaming source. If you use `getSource`, `getSourceWithFormat`, `createDataFrameFromOptions` or `create_data_frame_from_options`, you must specify these basic parameters using the connection options described here.

You can specify the connection options for Kinesis using the following arguments for the specified methods in the `GlueContext` class.

- Scala
  - `connectionOptions`: Use with `getSource`, `createDataFrameFromOptions`, `getSink`
  - `additionalOptions`: Use with `getCatalogSource`, `getCatalogSink`
  - `options`: Use with `getSourceWithFormat`, `getSinkWithFormat`
- Python
  - `connection_options`: Use with `create_data_frame_from_options`, `write_dynamic_frame_from_options`
  - `additional_options`: Use with `create_data_frame_from_catalog`, `write_dynamic_frame_from_catalog`
  - `options`: Use with `getSource`, `getSink`

For notes and restrictions about Streaming ETL jobs, consult the section called “Streaming ETL notes and restrictions”.

Configure Kinesis

To connect to a Kinesis data stream in an AWS Glue Spark job, you will need some prerequisites:

- If reading, the AWS Glue job must have Read access level IAM permissions to the Kinesis data stream.
- If writing, the AWS Glue job must have Write access level IAM permissions to the Kinesis data stream.

In certain cases, you will need to configure additional prerequisites:

- If your AWS Glue job is configured with **Additional network connections** (typically to connect to other datasets) and one of those connections provides Amazon VPC **Network options**, this will direct your job to communicate over Amazon VPC. In this case you will also need to configure your Kinesis data stream to communicate over Amazon VPC. You can do this by creating an interface VPC endpoint between your Amazon VPC and Kinesis data stream. For more information, see Using Kinesis Data Streams with Interface VPC Endpoints.
- When specifying Amazon Kinesis Data Streams in another account, you must setup the roles and policies to allow cross-account access. For more information, see Example: Read From a Kinesis Stream in a Different Account.

For more information about Streaming ETL job prerequisites, consult the section called “Streaming ETL jobs”.

Example: Reading from Kinesis streams

Example: Reading from Kinesis streams

Used in conjunction with the section called “forEachBatch”.

Example for Amazon Kinesis streaming source:

```
kinesis_options =
   { "streamARN": "arn:aws:kinesis:us-east-2:777788889999:stream/fromOptionsStream",
     "startingPosition": "TRIM_HORIZON",
     "inferSchema": "true",
     "classification": "json"
   }
data_frame_datasource0 =
 glueContext.create_data_frame.from_options(connection_type="kinesis",
 connection_options=kinesis_options)
```

Example: Writing to Kinesis streams

Example: Reading from Kinesis streams

Used in conjunction with the section called “forEachBatch”.

Example for Amazon Kinesis streaming source:

```
kinesis_options =
   { "streamARN": "arn:aws:kinesis:us-east-2:777788889999:stream/fromOptionsStream",
     "startingPosition": "TRIM_HORIZON",
     "inferSchema": "true",
     "classification": "json"
   }
data_frame_datasource0 =
 glueContext.create_data_frame.from_options(connection_type="kinesis",
 connection_options=kinesis_options)
```

Kinesis connection option reference

Designates connection options for Amazon Kinesis Data Streams.

Use the following connection options for Kinesis streaming data sources:

- `"streamARN"` (Required) Used for Read/Write. The ARN of the Kinesis data stream.
- `"classification"` (Required for read) Used for Read. The file format used by the data in the record. Required unless provided through the Data Catalog.
- `"streamName"` – (Optional) Used for Read. The name of a Kinesis data stream to read from. Used with `endpointUrl`.
- `"endpointUrl"` – (Optional) Used for Read. Default: "https://kinesis.us-east-1.amazonaws.com". The AWS endpoint of the Kinesis stream. You do not need to change this unless you are connecting to a special region.
- `"partitionKey"` – (Optional) Used for Write. The Kinesis partition key used when producing records.
- `"delimiter"` (Optional) Used for Read. The value separator used when `classification` is CSV. Default is "`,`."
- `"startingPosition"`: (Optional) Used for Read. The starting position in the Kinesis data stream to read data from. The possible values are `"latest"`, `"trim_horizon"`, `"earliest"`, or a Timestamp string in UTC format in the pattern `yyyy-mm-ddTHH:MM:SSZ` (where Z

represents a UTC timezone offset with a +/-. For example "2023-04-04T08:00:00-04:00"). The default value is `"latest"`. Note: the Timestamp string in UTC Format for `"startingPosition"` is supported only for AWS Glue version 4.0 or later.

- `"failOnDataLoss"`: (Optional) Fail the job if any active shard is missing or expired. The default value is `"false"`.
- `"awsSTSRoleARN"`: (Optional) Used for Read/Write. The Amazon Resource Name (ARN) of the role to assume using AWS Security Token Service (AWS STS). This role must have permissions for describe or read record operations for the Kinesis data stream. You must use this parameter when accessing a data stream in a different account. Used in conjunction with `"awsSTSSessionName"`.
- `"awsSTSSessionName"`: (Optional) Used for Read/Write. An identifier for the session assuming the role using AWS STS. You must use this parameter when accessing a data stream in a different account. Used in conjunction with `"awsSTSRoleARN"`.
- `"awsSTSEndpoint"`: (Optional) The AWS STS endpoint to use when connecting to Kinesis with an assumed role. This allows using the regional AWS STS endpoint in a VPC, which is not possible with the default global endpoint.
- `"maxFetchTimeInMs"`: (Optional) Used for Read. The maximum time spent for the job executor to read records for the current batch from the Kinesis data stream, specified in milliseconds (ms). Multiple `GetRecords` API calls may be made within this time. The default value is `1000`.
- `"maxFetchRecordsPerShard"`: (Optional) Used for Read. The maximum number of records to fetch per shard in the Kinesis data stream per microbatch. Note: The client can exceed this limit if the streaming job has already read extra records from Kinesis (in the same get-records call). If `maxFetchRecordsPerShard` needs to be strict then it needs to be a multiple of `maxRecordPerRead`. The default value is `100000`.
- `"maxRecordPerRead"`: (Optional) Used for Read. The maximum number of records to fetch from the Kinesis data stream in each `getRecords` operation. The default value is `10000`.
- `"addIdleTimeBetweenReads"`: (Optional) Used for Read. Adds a time delay between two consecutive `getRecords` operations. The default value is `"False"`. This option is only configurable for Glue version 2.0 and above.
- `"idleTimeBetweenReadsInMs"`: (Optional) Used for Read. The minimum time delay between two consecutive `getRecords` operations, specified in ms. The default value is `1000`. This option is only configurable for Glue version 2.0 and above.
- `"describeShardInterval"`: (Optional) Used for Read. The minimum time interval between two `ListShards` API calls for your script to consider resharding. For more information, see

Strategies for Resharding in *Amazon Kinesis Data Streams Developer Guide*. The default value is `1s`.

- `"numRetries"`: (Optional) Used for Read. The maximum number of retries for Kinesis Data Streams API requests. The default value is `3`.
- `"retryIntervalMs"`: (Optional) Used for Read. The cool-off time period (specified in ms) before retrying the Kinesis Data Streams API call. The default value is `1000`.
- `"maxRetryIntervalMs"`: (Optional) Used for Read. The maximum cool-off time period (specified in ms) between two retries of a Kinesis Data Streams API call. The default value is `10000`.
- `"avoidEmptyBatches"`: (Optional) Used for Read. Avoids creating an empty microbatch job by checking for unread data in the Kinesis data stream before the batch is started. The default value is `"False"`.
- `"schema"`: (Required when inferSchema set to false) Used for Read. The schema to use to process the payload. If classification is `avro` the provided schema must be in the Avro schema format. If the classification is not `avro` the provided schema must be in the DDL schema format.

  The following are schema examples.

  Example in DDL schema format

  ```
  `column1` INT, `column2` STRING , `column3` FLOAT
  ```

  Example in Avro schema format

  ```
  {
    "type":"array",
    "items":
    {
      "type":"record",
      "name":"test",
      "fields":
      [
        {
          "name":"_id",
          "type":"string"
        },
        {
          "name":"index",
          "type":
  ```

```
          [
            "int",
            "string",
            "float"
          ]
        }
      ]
    }
  }
```

- `"inferSchema"`: (Optional) Used for Read. The default value is 'false'. If set to 'true', the schema will be detected at runtime from the payload within `foreachbatch`.
- `"avroSchema"`: (Deprecated) Used for Read. Parameter used to specify a schema of Avro data when Avro format is used. This parameter is now deprecated. Use the `schema` parameter.
- `"addRecordTimestamp"`: (Optional) Used for Read. When this option is set to 'true', the data output will contain an additional column named "__src_timestamp" that indicates the time when the corresponding record received by the stream. The default value is 'false'. This option is supported in AWS Glue version 4.0 or later.
- `"emitConsumerLagMetrics"`: (Optional) Used for Read. When the option is set to 'true', for each batch, it will emit the metrics for the duration between the oldest record received by the stream and the time it arrives in AWS Glue to CloudWatch. The metric's name is "glue.driver.streaming.maxConsumerLagInMs". The default value is 'false'. This option is supported in AWS Glue version 4.0 or later.
- `"fanoutConsumerARN"`: (Optional) Used for Read. The ARN of a Kinesis stream consumer for the stream specified in `streamARN`. Used to enable enhanced fan-out mode for your Kinesis connection. For more information on consuming a Kinesis stream with enhanced fan-out, see the section called "Using enhanced fan-out in Kinesis streaming jobs".
- `"recordMaxBufferedTime"` – (Optional) Used for Write. Default: 1000 (ms). Maximum time a record is buffered while waiting to be written.
- `"aggregationEnabled"` – (Optional) Used for Write. Default: true. Specifies if records should be aggregated before sending them to Kinesis.
- `"aggregationMaxSize"` – (Optional) Used for Write. Default: 51200 (bytes). If a record is larger than this limit, it will bypass the aggregator. Note Kinesis enforces a limit of 50KB on record size. If you set this beyond 50KB, oversize records will be rejected by Kinesis.
- `"aggregationMaxCount"` – (Optional) Used for Write. Default: 4294967295. Maximum number of items to pack into an aggregated record.

- `"producerRateLimit"` – (Optional) Used for Write. Default: 150 (%). Limits per-shard throughput sent from a single producer (such as your job), as a percentage of the backend limit.
- `"collectionMaxCount"` – (Optional) Used for Write. Default: 500. Maximum number of items to pack into an PutRecords request.
- `"collectionMaxSize"` – (Optional) Used for Write. Default: 5242880 (bytes). Maximum amount of data to send with a PutRecords request.

Using enhanced fan-out in Kinesis streaming jobs

An enhanced fan-out consumer is able to recieve records from a Kinesis stream with dedicated throughput that can be greater than typical consumers. This is done by optimizing the transfer protocol used to provide data to a Kinesis consumer, such as your job. For more information about Kinesis Enhanced Fan-Out, see the Kinesis documentation.

In enhanced fan-out mode, the `maxRecordPerRead` and `idleTimeBetweenReadsInMs` connection options no longer apply, as those parameters are not configurable when using enhanced fan-out. The configuration options for retries perform as described.

Use the following procedures to enable and disable enhanced fan-out for your streaming job. You should register a stream consumer for each job that will consume data from your stream.

To enable enhanced fan-out consumption on your job:

1. Register a stream consumer for your job using the Kinesis API. Follow the instructions to *register a consumer with enhanced fan-out using the Kinesis Data Streams API* in the Kinesis documentation. You will only need to follow the first step - calling RegisterStreamConsumer. Your request should return an ARN, *`consumerARN`*.
2. Set the connection option `fanoutConsumerARN` to *`consumerARN`* in your connection method arguments.
3. Restart your job.

To disable enhanced fan-out consumption on your job:

1. Remove the `fanoutConsumerARN` connection option from your method call.
2. Restart your job.
3. Follow the instructions to *deregister a consumer* in the Kinesis documentation. These instructions apply to the console, but can also be achieved through the Kinesis API. For

more information about stream consumer deregistration through the Kinesis API, consult DeregisterStreamConsumer in the Kinesis documentation.

Amazon S3 connections

You can use AWS Glue for Spark to read and write files in Amazon S3. AWS Glue for Spark supports many common data formats stored in Amazon S3 out of the box, including CSV, Avro, JSON, Orc and Parquet. For more information about supported data formats, see the section called "Data format options". Each data format may support a different set of AWS Glue features. Consult the page for your data format for the specifics of feature support. Additionally, you can read and write versioned files stored in the Hudi, Iceberg and Delta Lake data lake frameworks. For more information about data lake frameworks, see the section called "Data lake frameworks".

With AWS Glue you can partition your Amazon S3 objects into a folder structure while writing, then retrieve it by partition to improve performance using simple configuration. You can also set configuration to group small files together when transforming your data to improve performance. You can read and write `bzip2` and `gzip` archives in Amazon S3.

Topics

- Configuring S3 connections
- Amazon S3 connection option reference
- Deprecated connection syntaxes for data formats
- Excluding Amazon S3 storage classes
- Managing partitions for ETL output in AWS Glue
- Reading input files in larger groups
- Amazon VPC endpoints for Amazon S3

Configuring S3 connections

To connect to Amazon S3 in a AWS Glue with Spark job, you will need some prerequisites:

- The AWS Glue job must have IAM permissions for relevant Amazon S3 buckets.

In certain cases, you will need to configure additional prerequisites:

- When configuring cross-account access, appropriate access controls on the Amazon S3 bucket.

- For security reasons, you may choose to route your Amazon S3 requests through an Amazon VPC. This approach can introduce bandwidth and availability challenges. For more information, see the section called “Amazon VPC endpoints for Amazon S3”.

Amazon S3 connection option reference

Designates a connection to Amazon S3.

Since Amazon S3 manages files rather than tables, in addition to specifying the connection properties provided in this document, you will need to specify additional configuration about your file type. You specify this information through data format options. For more information about format options, see the section called “Data format options”. You can also specify this information by integrating with the AWS Glue Data Catalog.

For an example of the distinction between connection options and format options, consider how the the section called “create_dynamic_frame_from_options” method takes `connection_type`, `connection_options`, `format` and `format_options`. This section specifically discusses parameters provided to `connection_options`.

Use the following connection options with `"connectionType": "s3"`:

- `"paths"`: (Required) A list of the Amazon S3 paths to read from.
- `"exclusions"`: (Optional) A string containing a JSON list of Unix-style glob patterns to exclude. For example, `"[\"**.pdf\"]"` excludes all PDF files. For more information about the glob syntax that AWS Glue supports, see Include and Exclude Patterns.
- `"compressionType"`: or `"compression"`: (Optional) Specifies how the data is compressed. Use `"compressionType"` for Amazon S3 sources and `"compression"` for Amazon S3 targets. This is generally not necessary if the data has a standard file extension. Possible values are `"gzip"` and `"bzip2"`). Additional compression formats may be supported for specific formats. For the specifics of feature support, consult the data format page.
- `"groupFiles"`: (Optional) Grouping files is turned on by default when the input contains more than 50,000 files. To turn on grouping with fewer than 50,000 files, set this parameter to `"inPartition"`. To disable grouping when there are more than 50,000 files, set this parameter to `"none"`.
- `"groupSize"`: (Optional) The target group size in bytes. The default is computed based on the input data size and the size of your cluster. When there are fewer than 50,000 input files, `"groupFiles"` must be set to `"inPartition"` for this to take effect.

- `"recurse"`: (Optional) If set to true, recursively reads files in all subdirectories under the specified paths.
- `"maxBand"`: (Optional, advanced) This option controls the duration in milliseconds after which the `s3` listing is likely to be consistent. Files with modification timestamps falling within the last `maxBand` milliseconds are tracked specially when using `JobBookmarks` to account for Amazon S3 eventual consistency. Most users don't need to set this option. The default is 900000 milliseconds, or 15 minutes.
- `"maxFilesInBand"`: (Optional, advanced) This option specifies the maximum number of files to save from the last `maxBand` seconds. If this number is exceeded, extra files are skipped and only processed in the next job run. Most users don't need to set this option.
- `"isFailFast"`: (Optional) This option determines if an AWS Glue ETL job throws reader parsing exceptions. If set to `true`, jobs fail fast if four retries of the Spark task fail to parse the data correctly.
- `"catalogPartitionPredicate"`: (Optional) Used for Read. The contents of a SQL `WHERE` clause. Used when reading from Data Catalog tables with a very large quantity of partitions. Retrieves matching partitions from Data Catalog indices. Used with `push_down_predicate`, an option on the the section called "create_dynamic_frame_from_catalog" method (and other similar methods). For more information, see the section called "Catalog partition predicates".
- `"partitionKeys"`: (Optional) Used for Write. An array of column label strings. AWS Glue will partition your data as specified by this configuration. For more information, see the section called "Writing partitions".
- `"excludeStorageClasses"`: (Optional) Used for Read. An array of strings specifying Amazon S3 storage classes. AWS Glue will exclude Amazon S3 objects based on this configuration. For more information, see the section called "Excluding Amazon S3 storage classes".

Deprecated connection syntaxes for data formats

Certain data formats can be accessed using a specific connection type syntax. This syntax is deprecated. We recommend you specify your formats using the `s3` connection type and the format options provided in the section called "Data format options" instead.

"connectionType": "Orc"

Designates a connection to files stored in Amazon S3 in the Apache Hive Optimized Row Columnar (ORC) file format.

Use the following connection options with `"connectionType": "orc"`:

- `paths`: (Required) A list of the Amazon S3 paths to read from.
- *(Other option name/value pairs)*: Any additional options, including formatting options, are passed directly to the SparkSQL `DataSource`.

"connectionType": "parquet"

Designates a connection to files stored in Amazon S3 in the Apache Parquet file format.

Use the following connection options with `"connectionType": "parquet"`:

- `paths`: (Required) A list of the Amazon S3 paths to read from.
- *(Other option name/value pairs)*: Any additional options, including formatting options, are passed directly to the SparkSQL `DataSource`.

Excluding Amazon S3 storage classes

If you're running AWS Glue ETL jobs that read files or partitions from Amazon Simple Storage Service (Amazon S3), you can exclude some Amazon S3 storage class types.

The following storage classes are available in Amazon S3:

- `STANDARD` — For general-purpose storage of frequently accessed data.
- `INTELLIGENT_TIERING` — For data with unknown or changing access patterns.
- `STANDARD_IA` and `ONEZONE_IA` — For long-lived, but less frequently accessed data.
- `GLACIER`, `DEEP_ARCHIVE`, and `REDUCED_REDUNDANCY` — For long-term archive and digital preservation.

For more information, see Amazon S3 Storage Classes in the *Amazon S3 Developer Guide*.

The examples in this section show how to exclude the `GLACIER` and `DEEP_ARCHIVE` storage classes. These classes allow you to list files, but they won't let you read the files unless they are restored. (For more information, see Restoring Archived Objects in the *Amazon S3 Developer Guide*.)

By using storage class exclusions, you can ensure that your AWS Glue jobs will work on tables that have partitions across these storage class tiers. Without exclusions, jobs that read data from these tiers fail with the following error: AmazonS3Exception: The operation is not valid for the object's storage class.

There are different ways that you can filter Amazon S3 storage classes in AWS Glue.

Topics

- Excluding Amazon S3 storage classes when creating a Dynamic Frame
- Excluding Amazon S3 storage classes on a Data Catalog table

Excluding Amazon S3 storage classes when creating a Dynamic Frame

To exclude Amazon S3 storage classes while creating a dynamic frame, use `excludeStorageClasses` in `additionalOptions`. AWS Glue automatically uses its own Amazon S3 `Lister` implementation to list and exclude files corresponding to the specified storage classes.

The following Python and Scala examples show how to exclude the `GLACIER` and `DEEP_ARCHIVE` storage classes when creating a dynamic frame.

Python example:

```
glueContext.create_dynamic_frame.from_catalog(
    database = "my_database",
    tableName = "my_table_name",
    redshift_tmp_dir = "",
    transformation_ctx = "my_transformation_context",
    additional_options = {
        "excludeStorageClasses" : ["GLACIER", "DEEP_ARCHIVE"]
    }
)
```

Scala example:

```
val* *df = glueContext.getCatalogSource(
    nameSpace, tableName, "", "my_transformation_context",
    additionalOptions = JsonOptions(
        Map("excludeStorageClasses" -> List("GLACIER", "DEEP_ARCHIVE"))
    )
).getDynamicFrame()
```

Excluding Amazon S3 storage classes on a Data Catalog table

You can specify storage class exclusions to be used by an AWS Glue ETL job as a table parameter in the AWS Glue Data Catalog. You can include this parameter in the `CreateTable` operation using the AWS Command Line Interface (AWS CLI) or programmatically using the API. For more information, see Table Structure and CreateTable.

You can also specify excluded storage classes on the AWS Glue console.

To exclude Amazon S3 storage classes (console)

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. In the navigation pane on the left, choose **Tables**.
3. Choose the table name in the list, and then choose **Edit table**.
4. In **Table properties**, add **`excludeStorageClasses`** as a key and **`[\"GLACIER\", \"DEEP_ARCHIVE\"]`** as a value.
5. Choose **Apply**.

Managing partitions for ETL output in AWS Glue

Partitioning is an important technique for organizing datasets so they can be queried efficiently. It organizes data in a hierarchical directory structure based on the distinct values of one or more columns.

For example, you might decide to partition your application logs in Amazon Simple Storage Service (Amazon S3) by date, broken down by year, month, and day. Files that correspond to a single day's worth of data are then placed under a prefix such as `s3://my_bucket/logs/year=2018/month=01/day=23/`. Systems like Amazon Athena, Amazon Redshift Spectrum, and now AWS Glue can use these partitions to filter data by partition value without having to read all the underlying data from Amazon S3.

Crawlers not only infer file types and schemas, they also automatically identify the partition structure of your dataset when they populate the AWS Glue Data Catalog. The resulting partition columns are available for querying in AWS Glue ETL jobs or query engines like Amazon Athena.

After you crawl a table, you can view the partitions that the crawler created. In the AWS Glue console, choose **Tables** in the left navigation pane. Choose the table created by the crawler, and then choose **View Partitions**.

For Apache Hive-style partitioned paths in `key=val` style, crawlers automatically populate the column name using the key name. Otherwise, it uses default names like `partition_0`, `partition_1`, and so on. You can change the default names on the console. To do so, navigate to the table. Check if indexes exist under the **Indexes** tab. If that's the case, you need to delete them to proceed (you can recreate them using the new column names afterwards). Then, choose **Edit Schema**, and modify the names of the partition columns there.

In your ETL scripts, you can then filter on the partition columns. Because the partition information is stored in the Data Catalog, use the `from_catalog` API calls to include the partition columns in the `DynamicFrame`. For example, use `create_dynamic_frame.from_catalog` instead of `create_dynamic_frame.from_options`.

Partitioning is an optimization technique that reduces data scan. For more information about the process of identifying when this technique is appropriate, consult Reduce the amount of data scan in the *Best practices for performance tuning AWS Glue for Apache Spark jobs* guide on AWS Prescriptive Guidance.

Pre-filtering using pushdown predicates

In many cases, you can use a pushdown predicate to filter on partitions without having to list and read all the files in your dataset. Instead of reading the entire dataset and then filtering in a DynamicFrame, you can apply the filter directly on the partition metadata in the Data Catalog. Then you only list and read what you actually need into a DynamicFrame.

For example, in Python, you could write the following.

```
glue_context.create_dynamic_frame.from_catalog(
    database = "my_S3_data_set",
    table_name = "catalog_data_table",
    push_down_predicate = my_partition_predicate)
```

This creates a DynamicFrame that loads only the partitions in the Data Catalog that satisfy the predicate expression. Depending on how small a subset of your data you are loading, this can save a great deal of processing time.

The predicate expression can be any Boolean expression supported by Spark SQL. Anything you could put in a `WHERE` clause in a Spark SQL query will work. For example, the predicate expression `pushDownPredicate = "(year=='2017' and month=='04')"` loads only the partitions in the Data Catalog that have both `year` equal to 2017 and `month` equal to 04. For more information, see the Apache Spark SQL documentation, and in particular, the Scala SQL functions reference.

Server-side filtering using catalog partition predicates

The `push_down_predicate` option is applied after listing all the partitions from the catalog and before listing files from Amazon S3 for those partitions. If you have a lot of partitions for a table, catalog partition listing can still incur additional time overhead. To address this overhead, you can use server-side partition pruning with the `catalogPartitionPredicate` option that uses partition indexes in the AWS Glue Data Catalog. This makes partition filtering much faster when you have millions of partitions in one table. You can use both `push_down_predicate` and `catalogPartitionPredicate` in `additional_options` together if your `catalogPartitionPredicate` requires predicate syntax that is not yet supported with the catalog partition indexes.

Python:

```
dynamic_frame = glueContext.create_dynamic_frame.from_catalog(
    database=dbname,
    table_name=tablename,
    transformation_ctx="datasource0",
    push_down_predicate="day>=10 and customer_id like '10%'",
    additional_options={"catalogPartitionPredicate":"year='2021' and month='06'"}
)
```

Scala:

```
val dynamicFrame = glueContext.getCatalogSource(
    database = dbname,
    tableName = tablename,
    transformationContext = "datasource0",
    pushDownPredicate="day>=10 and customer_id like '10%'",
    additionalOptions = JsonOptions("""{
        "catalogPartitionPredicate": "year='2021' and month='06'"}""")
    ).getDynamicFrame()
```

Note

The `push_down_predicate` and `catalogPartitionPredicate` use different syntaxes. The former one uses Spark SQL standard syntax and the later one uses JSQL parser.

Writing partitions

By default, a DynamicFrame is not partitioned when it is written. All of the output files are written at the top level of the specified output path. Until recently, the only way to write a DynamicFrame into partitions was to convert it to a Spark SQL DataFrame before writing.

However, DynamicFrames now support native partitioning using a sequence of keys, using the `partitionKeys` option when you create a sink. For example, the following Python code writes out a dataset to Amazon S3 in the Parquet format, into directories partitioned by the type field. From there, you can process these partitions using other systems, such as Amazon Athena.

```
glue_context.write_dynamic_frame.from_options(
    frame = projectedEvents,
    connection_type = "s3",
    connection_options = {"path": "$outpath", "partitionKeys": ["type"]},
    format = "parquet")
```

Reading input files in larger groups

You can set properties of your tables to enable an AWS Glue ETL job to group files when they are read from an Amazon S3 data store. These properties enable each ETL task to read a group of input files into a single in-memory partition, this is especially useful when there is a large number of small files in your Amazon S3 data store. When you set certain properties, you instruct AWS Glue to group files within an Amazon S3 data partition and set the size of the groups to be read. You can also set these options when reading from an Amazon S3 data store with the `create_dynamic_frame.from_options` method.

To enable grouping files for a table, you set key-value pairs in the parameters field of your table structure. Use JSON notation to set a value for the parameter field of your table. For more information about editing the properties of a table, see Viewing and managing table details.

You can use this method to enable grouping for tables in the Data Catalog with Amazon S3 data stores.

groupFiles

Set **groupFiles** to `inPartition` to enable the grouping of files within an Amazon S3 data partition. AWS Glue automatically enables grouping if there are more than 50,000 input files, as in the following example.

```
'groupFiles': 'inPartition'
```

groupSize

Set **groupSize** to the target size of groups in bytes. The **groupSize** property is optional, if not provided, AWS Glue calculates a size to use all the CPU cores in the cluster while still reducing the overall number of ETL tasks and in-memory partitions.

For example, the following sets the group size to 1 MB.

```
'groupSize': '1048576'
```

Note that the `groupsize` should be set with the result of a calculation. For example 1024 * 1024 = 1048576.

recurse

Set **recurse** to `True` to recursively read files in all subdirectories when specifying `paths` as an array of paths. You do not need to set **recurse** if `paths` is an array of object keys in Amazon S3, or if the input format is parquet/orc, as in the following example.

```
'recurse':True
```

If you are reading from Amazon S3 directly using the `create_dynamic_frame.from_options` method, add these connection options. For example, the following attempts to group files into 1 MB groups.

```
df = glueContext.create_dynamic_frame.from_options("s3", {'paths': ["s3://s3path/"],
 'recurse':True, 'groupFiles': 'inPartition', 'groupSize': '1048576'}, format="json")
```

> **Note**
>
> `groupFiles` is supported for DynamicFrames created from the following data formats: csv, ion, grokLog, json, and xml. This option is not supported for avro, parquet, and orc.

Amazon VPC endpoints for Amazon S3

For security reasons, many AWS customers run their applications within an Amazon Virtual Private Cloud environment (Amazon VPC). With Amazon VPC, you can launch Amazon EC2 instances into a virtual private cloud, which is logically isolated from other networks—including the public internet. With an Amazon VPC, you have control over its IP address range, subnets, routing tables, network gateways, and security settings.

> **Note**
>
> If you created your AWS account after 2013-12-04, you already have a default VPC in each AWS Region. You can immediately start using your default VPC without any additional configuration.
> For more information, see Your Default VPC and Subnets in the Amazon VPC User Guide.

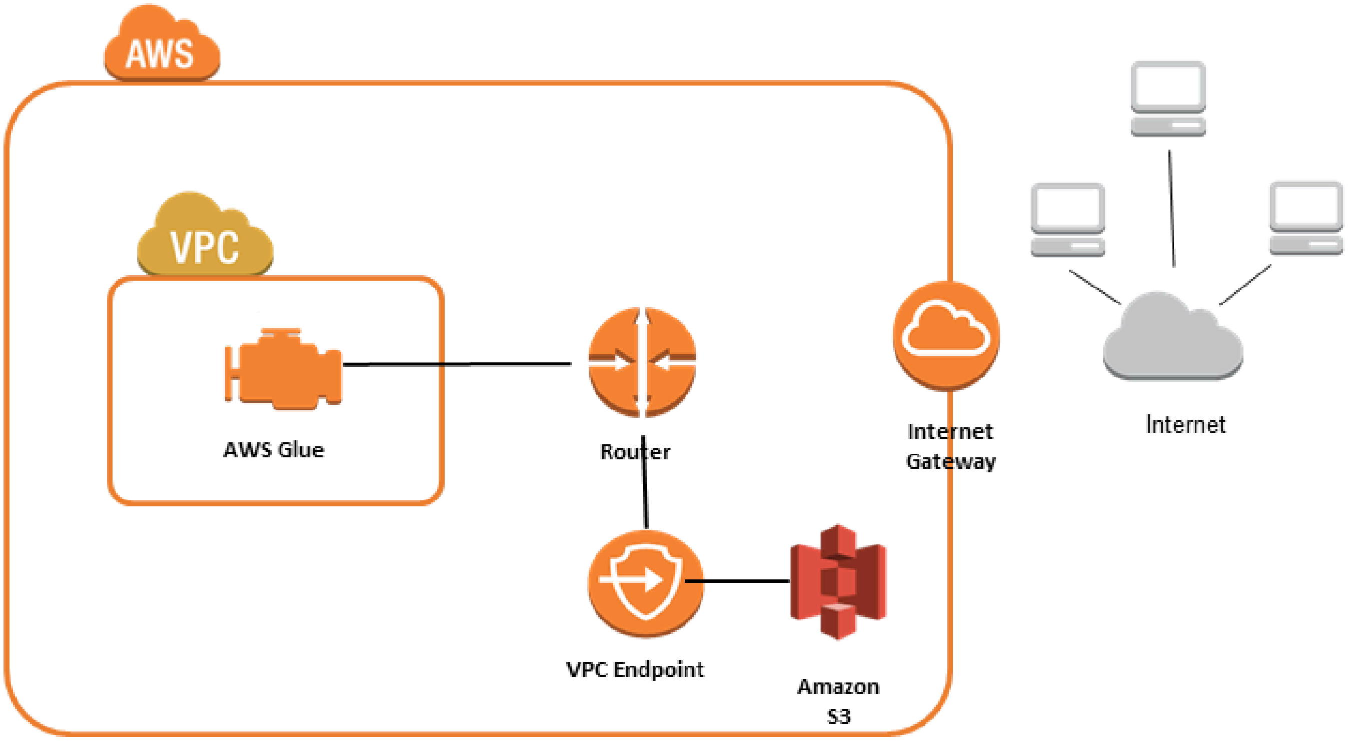
Many customers have legitimate privacy and security concerns about sending and receiving data across the public internet. Customers can address these concerns by using a virtual private network (VPN) to route all Amazon S3 network traffic through their own corporate network infrastructure. However, this approach can introduce bandwidth and availability challenges.

VPC endpoints for Amazon S3 can alleviate these challenges. A VPC endpoint for Amazon S3 enables AWS Glue to use private IP addresses to access Amazon S3 with no exposure to the public internet. AWS Glue does not require public IP addresses, and you don't need an internet gateway, a NAT device, or a virtual private gateway in your VPC. You use endpoint policies to control access to Amazon S3. Traffic between your VPC and the AWS service does not leave the Amazon network.

When you create a VPC endpoint for Amazon S3, any requests to an Amazon S3 endpoint within the Region (for example, *s3.us-west-2.amazonaws.com*) are routed to a private Amazon S3 endpoint within the Amazon network. You don't need to modify your applications running on Amazon EC2 instances in your VPC—the endpoint name remains the same, but the route to Amazon S3 stays entirely within the Amazon network, and does not access the public internet.

For more information about VPC endpoints, see VPC Endpoints in the Amazon VPC User Guide.

The following diagram shows how AWS Glue can use a VPC endpoint to access Amazon S3.


AWS
VPC
AWS Glue
Router
Internet Gateway
Internet
VPC Endpoint
Amazon S3

To set up access for Amazon S3

1. Sign in to the AWS Management Console and open the Amazon VPC console at https://console.aws.amazon.com/vpc/.
2. In the left navigation pane, choose **Endpoints**.
3. Choose **Create Endpoint**, and follow the steps to create an Amazon S3 VPC endpoint of type Gateway.

Amazon DocumentDB connections

You can use AWS Glue for Spark to read from and write to tables in Amazon DocumentDB. You can connect to Amazon DocumentDB using credentials stored in AWS Secrets Manager through a AWS Glue connection.

For more information about Amazon DocumentDB, consult the Amazon DocumentDB documentation.

> **Note**
>
> Amazon DocumentDB elastic clusters are not currently supported when using the AWS Glue connector. For more information about elastic clusters, see Using Amazon DocumentDB elastic clusters.

Reading and writing to Amazon DocumentDB collections

> **Note**
>
> When you create an ETL job that connects to Amazon DocumentDB, for the `Connections` job property, you must designate a connection object that specifies the virtual private cloud (VPC) in which Amazon DocumentDB is running. For the connection object, the connection type must be `JDBC`, and the `JDBC URL` must be `mongo://<DocumentDB_host>:27017`.

> **Note**
>
> These code samples were developed for AWS Glue 3.0. To migrate to AWS Glue 4.0, consult the section called "MongoDB". The `uri` parameter has changed.

> **Note**
>
> When using Amazon DocumentDB, `retryWrites` must be set to false in certain situations, such as when the document written specifies `_id`. For more information, consult Functional Differences with MongoDB in the Amazon DocumentDB documentation.

The following Python script demonstrates using connection types and connection options for reading and writing to Amazon DocumentDB.

```
import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext, SparkConf
from awsglue.context import GlueContext
from awsglue.job import Job
import time

## @params: [JOB_NAME]
args = getResolvedOptions(sys.argv, ['JOB_NAME'])

sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session

job = Job(glueContext)
job.init(args['JOB_NAME'], args)

output_path = "s3://some_bucket/output/" + str(time.time()) + "/"
documentdb_uri = "mongodb://<mongo-instanced-ip-address>:27017"
documentdb_write_uri = "mongodb://<mongo-instanced-ip-address>:27017"

read_docdb_options = {
    "uri": documentdb_uri,
    "database": "test",
    "collection": "coll",
    "username": "username",
    "password": "1234567890",
    "ssl": "true",
    "ssl.domain_match": "false",
    "partitioner": "MongoSamplePartitioner",
    "partitionerOptions.partitionSizeMB": "10",
    "partitionerOptions.partitionKey": "_id"
}

write_documentdb_options = {
    "retryWrites": "false",
    "uri": documentdb_write_uri,
    "database": "test",
    "collection": "coll",
    "username": "username",
    "password": "pwd"
}

# Get DynamicFrame from  DocumentDB
dynamic_frame2 =
 glueContext.create_dynamic_frame.from_options(connection_type="documentdb",

 connection_options=read_docdb_options)

# Write DynamicFrame to MongoDB and DocumentDB
glueContext.write_dynamic_frame.from_options(dynamic_frame2,
 connection_type="documentdb",

 connection_options=write_documentdb_options)

job.commit()
```

The following Scala script demonstrates using connection types and connection options for reading and writing to Amazon DocumentDB.

```
import com.amazonaws.services.glue.GlueContext
import com.amazonaws.services.glue.MappingSpec
import com.amazonaws.services.glue.errors.CallSite
import com.amazonaws.services.glue.util.GlueArgParser
import com.amazonaws.services.glue.util.Job
import com.amazonaws.services.glue.util.JsonOptions
import com.amazonaws.services.glue.DynamicFrame
import org.apache.spark.SparkContext
import scala.collection.JavaConverters._

object GlueApp {
  val DOC_URI: String = "mongodb://<mongo-instanced-ip-address>:27017"
  val DOC_WRITE_URI: String = "mongodb://<mongo-instanced-ip-address>:27017"
  lazy val documentDBJsonOption = jsonOptions(DOC_URI)
  lazy val writeDocumentDBJsonOption = jsonOptions(DOC_WRITE_URI)
  def main(sysArgs: Array[String]): Unit = {
    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)
    val args = GlueArgParser.getResolvedOptions(sysArgs, Seq("JOB_NAME").toArray)
    Job.init(args("JOB_NAME"), glueContext, args.asJava)

    // Get DynamicFrame from DocumentDB
    val resultFrame2: DynamicFrame = glueContext.getSource("documentdb",
 documentDBJsonOption).getDynamicFrame()

    // Write DynamicFrame to DocumentDB
    glueContext.getSink("documentdb", writeJsonOption).writeDynamicFrame(resultFrame2)

    Job.commit()
  }

  private def jsonOptions(uri: String): JsonOptions = {
    new JsonOptions(
      s"""{"uri": "${uri}",
         |"database":"test",
         |"collection":"coll",
         |"username": "username",
         |"password": "pwd",
         |"ssl":"true",
         |"ssl.domain_match":"false",
         |"partitioner": "MongoSamplePartitioner",
         |"partitionerOptions.partitionSizeMB": "10",
         |"partitionerOptions.partitionKey": "_id"}""".stripMargin)
  }
}
```

Amazon DocumentDB connection option reference

Designates a connection to Amazon DocumentDB (with MongoDB compatibility).

Connection options differ for a source connection and a sink connection.

"connectionType": "Documentdb" as source

Use the following connection options with `"connectionType": "documentdb"` as a source:

- `"uri"`: (Required) The Amazon DocumentDB host to read from, formatted as `mongodb://<host>:<port>`.
- `"database"`: (Required) The Amazon DocumentDB database to read from.
- `"collection"`: (Required) The Amazon DocumentDB collection to read from.
- `"username"`: (Required) The Amazon DocumentDB user name.
- `"password"`: (Required) The Amazon DocumentDB password.

- `"ssl"`: (Required if using SSL) If your connection uses SSL, then you must include this option with the value `"true"`.
- `"ssl.domain_match"`: (Required if using SSL) If your connection uses SSL, then you must include this option with the value `"false"`.
- `"batchSize"`: (Optional): The number of documents to return per batch, used within the cursor of internal batches.
- `"partitioner"`: (Optional): The class name of the partitioner for reading input data from Amazon DocumentDB. The connector provides the following partitioners:
  - `MongoDefaultPartitioner` (default) (Not supported in AWS Glue 4.0)
  - `MongoSamplePartitioner` (Not supported in AWS Glue 4.0)
  - `MongoShardedPartitioner`
  - `MongoSplitVectorPartitioner`
  - `MongoPaginateByCountPartitioner`
  - `MongoPaginateBySizePartitioner` (Not supported in AWS Glue 4.0)
- `"partitionerOptions"` (Optional): Options for the designated partitioner. The following options are supported for each partitioner:
  - `MongoSamplePartitioner`: `partitionKey`, `partitionSizeMB`, `samplesPerPartition`
  - `MongoShardedPartitioner`: `shardkey`
  - `MongoSplitVectorPartitioner`: `partitionKey`, partitionSizeMB
  - `MongoPaginateByCountPartitioner`: `partitionKey`, `numberOfPartitions`
  - `MongoPaginateBySizePartitioner`: `partitionKey`, partitionSizeMB

  For more information about these options, see Partitioner Configuration in the MongoDB documentation.

"connectionType": "Documentdb" as sink

Use the following connection options with `"connectionType": "documentdb"` as a sink:

- `"uri"`: (Required) The Amazon DocumentDB host to write to, formatted as `mongodb://<host>:<port>`.
- `"database"`: (Required) The Amazon DocumentDB database to write to.
- `"collection"`: (Required) The Amazon DocumentDB collection to write to.

- `"username"`: (Required) The Amazon DocumentDB user name.
- `"password"`: (Required) The Amazon DocumentDB password.
- `"extendedBsonTypes"`: (Optional) If `true`, allows extended BSON types when writing data to Amazon DocumentDB. The default is `true`.
- `"replaceDocument"`: (Optional) If `true`, replaces the whole document when saving datasets that contain an `_id` field. If `false`, only fields in the document that match the fields in the dataset are updated. The default is `true`.
- `"maxBatchSize"`: (Optional): The maximum batch size for bulk operations when saving data. The default is 512.
- `"retryWrites"`: (Optional): Automatically retry certain write operations a single time if AWS Glue encounters a network error.

OpenSearch Service connections

You can use AWS Glue for Spark to read from and write to tables in OpenSearch Service in AWS Glue 4.0 and later versions. You can define what to read from OpenSearch Service with an OpenSearch query. You connect to OpenSearch Service using HTTP basic authentication credentials stored in AWS Secrets Manager through a AWS Glue connection. This feature is not compatible with OpenSearch Service serverless.

For more information about Amazon OpenSearch Service, see the [Amazon OpenSearch Service documentation].

Configuring OpenSearch Service connections

To connect to OpenSearch Service from AWS Glue, you will need to create and store your OpenSearch Service credentials in a AWS Secrets Manager secret, then associate that secret with a OpenSearch Service AWS Glue connection.

Prerequisites:

- Identify the domain endpoint, *`aosEndpoint`* and port, *`aosPort`* you would like to read from, or create the resource by following instructions in the Amazon OpenSearch Service documentation. For more information on creating a domain, see [Creating and managing Amazon OpenSearch Service domains] in the Amazon OpenSearch Service documentation.

  An Amazon OpenSearch Service domain endpoint will have the following default form, https://search-*`domainName`*-*`unstructuredIdContent`*.*`region`*.es.amazonaws.com. For more information on identifying your domain endpoint, see Creating and managing Amazon OpenSearch Service domains in the Amazon OpenSearch Service documentation.

Identify or generate HTTP basic authentication credentials, *`aosUser`* and *`aosPassword`* for your domain.

To configure a connection to OpenSearch Service:

1. In AWS Secrets Manager, create a secret using your OpenSearch Service credentials. To create a secret in Secrets Manager, follow the tutorial available in Create an AWS Secrets Manager secret in the AWS Secrets Manager documentation. After creating the secret, keep the Secret name, *`secretName`* for the next step.
    - When selecting **Key/value pairs**, create a pair for the key `USERNAME` with the value *`aosUser`*.
    - When selecting **Key/value pairs**, create a pair for the key `PASSWORD` with the value *`aosPassword`*.
2. In the AWS Glue console, create a connection by following the steps in the section called "Adding an AWS Glue connection". After creating the connection, keep the connection name, *`connectionName`*, for future use in AWS Glue.
    - When selecting a **Connection type**, select OpenSearch Service.
    - When selecting a Domain endpoint, provide *`aosEndpoint`*.
    - When selecting a port, provide *`aosPort`*.
    - When selecting an **AWS Secret**, provide *`secretName`*.

After creating a AWS Glue OpenSearch Service connection, you will need to perform the following steps before running your AWS Glue job:

- Grant the IAM role associated with your AWS Glue job permission to read *`secretName`*.
- In your AWS Glue job configuration, provide *`connectionName`* as an **Additional network connection**.

Reading from OpenSearch Service indexes

Prerequisites:

- A OpenSearch Service index you would like to read from, *aosIndex*.
- A AWS Glue OpenSearch Service connection configured to provide auth and network location information. To acquire this, complete the steps in the previous procedure, *To configure a connection to OpenSearch Service*. You will need the name of the AWS Glue connection, *connectionName*.

This example reads an index from Amazon OpenSearch Service. You will need to provide the pushdown parameter.

For example:

```
opensearch_read = glueContext.create_dynamic_frame.from_options(
    connection_type="opensearch",
    connection_options={
        "connectionName": "connectionName",
        "opensearch.resource": "aosIndex",
        "pushdown": "true",
    }
)
```

You can also provide a query string to filter the results returned in your DynamicFrame. You will need to configure `opensearch.query`.

`opensearch.query` can take a URL query parameter string *queryString* or a query DSL JSON object *queryObject*. For more information about the query DSL, see Query DSL in the OpenSearch documentation. To provide a URL query parameter string, prepend ?q= to your query, as you would in a fully qualified URL. To provide a query DSL object, string escape the JSON object before providing it.

For example:

```

            queryObject = "{ "query": { "multi_match": { "query": "Sample", "fields":
 [ "sample" ] } } }"
            queryString = "?q=queryString"

            opensearch_read_query = glueContext.create_dynamic_frame.from_options(
    connection_type="opensearch",
    connection_options={
        "connectionName": "connectionName",
        "opensearch.resource": "aosIndex",
        "opensearch.query": queryString,
        "pushdown": "true",
    }
)
```

For more information about how to build a query outside of its specific syntax, see [Query string syntax]() in the OpenSearch documentation.

When reading from OpenSearch collections that contain array type data, you must specify which fields are array type in your method call using the `opensearch.read.field.as.array.include` parameter.

For example, when reading the following document, you will encounter the `genre` and `actor` array fields:

```
{
    "_index": "movies",
    "_id": "2",
    "_version": 1,
    "_seq_no": 0,
    "_primary_term": 1,
    "found": true,
    "_source": {
        "director": "Frankenheimer, John",
        "genre": [
            "Drama",
            "Mystery",
            "Thriller",
            "Crime"
        ],
        "year": 1962,
        "actor": [
            "Lansbury, Angela",
            "Sinatra, Frank",
            "Leigh, Janet",
            "Harvey, Laurence",
            "Silva, Henry",
            "Frees, Paul",
            "Gregory, James",
            "Bissell, Whit",
            "McGiver, John",
                "Parrish, Leslie",
                "Edwards, James",
                "Flowers, Bess",
                "Dhiegh, Khigh",
                "Payne, Julie",
                "Kleeb, Helen",
                "Gray, Joe",
                "Nalder, Reggie",
                "Stevens, Bert",
                "Masters, Michael",
                "Lowell, Tom"
            ],
            "title": "The Manchurian Candidate"
        }
    }
```

In this case, you would include those field names in your method call. For example:

```
"opensearch.read.field.as.array.include": "genre,actor"
```

If your array field is nested inside of your document structure, refer to it using dot notation: `"genre,actor,foo.bar.baz"`. This would specify an array `baz` included in your source document through the embedded document `foo` containing the embedded document `bar`.

Writing to OpenSearch Service tables

This example writes information from an existing DynamicFrame, *`dynamicFrame`* to OpenSearch Service. If the index already has information, AWS Glue will append data from your DynamicFrame. You will need to provide the `pushdown` parameter.

Prerequisites:

- A OpenSearch Service table you would like to write to. You will need identification information for the table. Let's call this *`tableName`*.
- A AWS Glue OpenSearch Service connection configured to provide auth and network location information. To acquire this, complete the steps in the previous procedure, *To configure a connection to OpenSearch Service*. You will need the name of the AWS Glue connection, *`connectionName`*.

For example:

```
glueContext.write_dynamic_frame.from_options(
    frame=dynamicFrame,
    connection_type="opensearch",
    connection_options={
      "connectionName": "connectionName",
      "opensearch.resource": "aosIndex",
    },
)
```

OpenSearch Service connection option reference

- `connectionName` — Required. Used for Read/Write. The name of a AWS Glue OpenSearch Service connection configured to provide auth and network location information to your connection method.
- `opensearch.resource` — Required. Used for Read/Write. Valid Values: OpenSearch index names. The name of the index your connection method will interact with.
- `opensearch.query` — Used for Read. Valid Values: String escaped JSON or, when this string begins with `?`, the search part of a URL. An OpenSearch query that filters what should be retrieved when reading. For more information on using this parameter, consult the previous section the section called "Read from OpenSearch Service".
- `pushdown` — Required if. Used for Read. Valid Values: boolean. Instructs Spark to pass read queries down to OpenSearch so the database only returns relevant documents.
- `opensearch.read.field.as.array.include` — Required if reading array type data. Used for Read. Valid Values: comma separated lists of field names. Specifies fields to read as arrays from OpenSearch documents. For more information on using this parameter, consult the previous section the section called "Read from OpenSearch Service".

Redshift connections

You can use AWS Glue for Spark to read from and write to tables in Amazon Redshift databases. When connecting to Amazon Redshift databases, AWS Glue moves data through Amazon S3 to achieve maximum throughput, using the Amazon Redshift SQL `COPY` and `UNLOAD` commands. In AWS Glue 4.0 and later, you can use the Amazon Redshift integration for Apache Spark to read and write with optimizations and features specific to Amazon Redshift beyond those available when connecting through previous versions.

Learn about how AWS Glue is making it easier than ever for Amazon Redshift users to migrate to AWS Glue for serverless data integration and ETL.

Configuring Redshift connections

To use Amazon Redshift clusters in AWS Glue, you will need some prerequisites:

- An Amazon S3 directory to use for temporary storage when reading from and writing to the database.
- An Amazon VPC enabling communication between your Amazon Redshift cluster, your AWS Glue job and your Amazon S3 directory.
- Appropriate IAM permissions on the AWS Glue job and Amazon Redshift cluster.

Configuring IAM roles

Set up the role for the Amazon Redshift cluster

Your Amazon Redshift cluster needs to be able to read and write to Amazon S3 in order to integrate with AWS Glue jobs. To allow this, you can associate IAM roles with the Amazon Redshift cluster you want to connect to. Your role should have a policy allowing read from and write to your Amazon S3 temporary directory. Your role should have a trust relationship allowing the `redshift.amazonaws.com` service to `AssumeRole`.

To associate an IAM role with Amazon Redshift

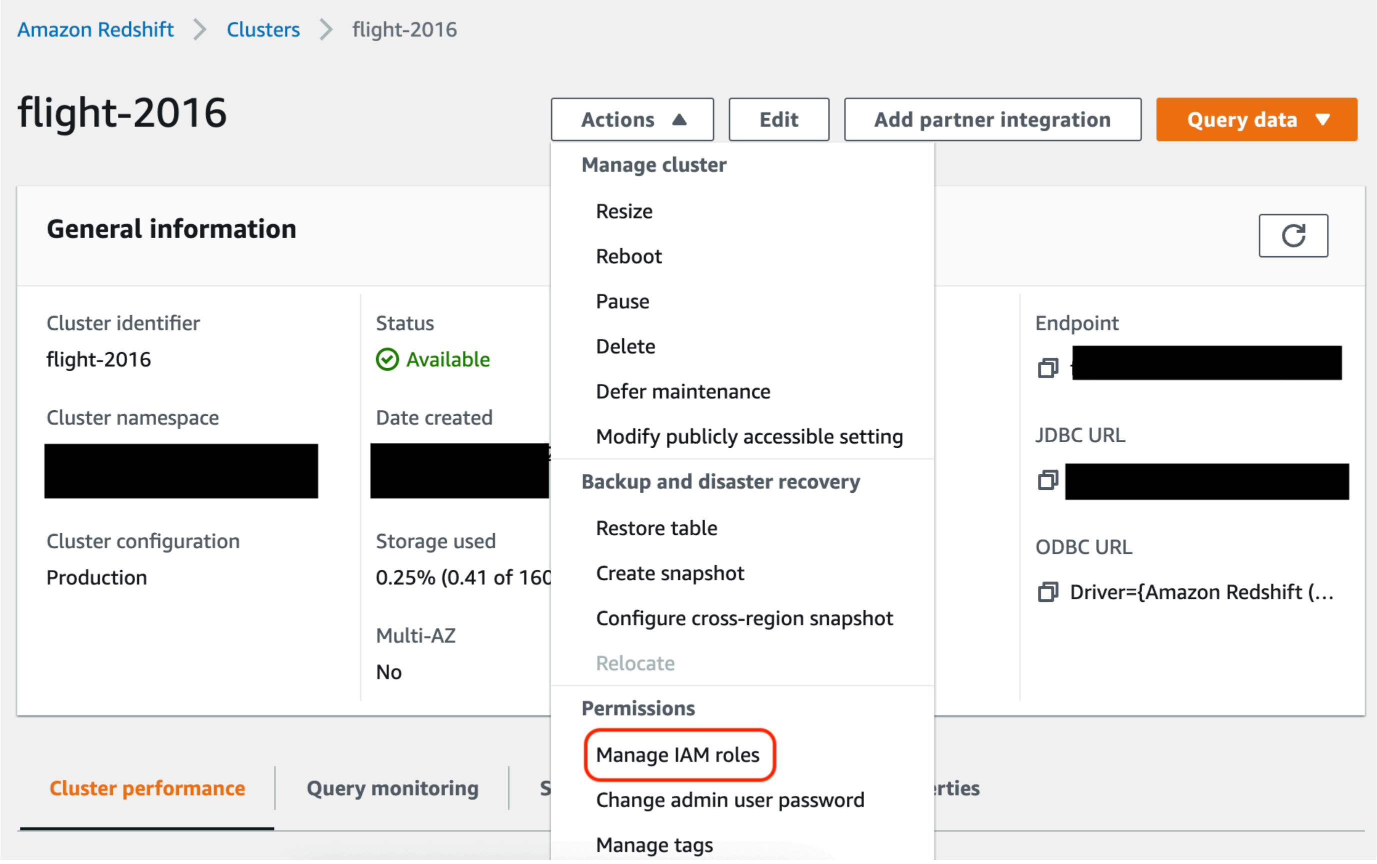
1. **Prerequisites:** An Amazon S3 bucket or directory used for the temporary storage of files.
2. Identify which Amazon S3 permissions your Amazon Redshift cluster will need. When moving data to and from an Amazon Redshift cluster, AWS Glue jobs issue COPY and UNLOAD statements against Amazon Redshift. If your job modifies a table in Amazon Redshift, AWS Glue will also issue CREATE LIBRARY statements. For information on specific Amazon S3 permissions required for Amazon Redshift to execute these statements, refer to the Amazon Redshift documentation: Amazon Redshift: Permissions to access other AWS Resources.
3. In the IAM console, create an IAM policy with the necessary permissions. For more information about creating a policy Creating IAM policies.
4. In the IAM console, create a role and trust relationship allowing Amazon Redshift to assume the role. Follow the instructions in the IAM documentation To create a role for an AWS service (console)

- When asked to choose an AWS service use case, choose "Redshift - Customizable".
- When asked to attach a policy, choose the policy you previously defined.

> **Note**
>
> For more information about configuring roles for Amazon Redshift, see Authorizing Amazon Redshift to access other AWS services on your behalf in the Amazon Redshift documentation.

5. In the Amazon Redshift console, associate the role with your Amazon Redshift cluster. Follow the instructions in the Amazon Redshift documentation.

   Select the highlighted option in the Amazon Redshift console to configure this setting:



Amazon Redshift > Clusters > flight-2016
flight-2016
Actions
Edit
Add partner integration
Query data
Manage cluster
Resize
Reboot
Pause
Delete
Defer maintenance
Modify publicly accessible setting
Backup and disaster recovery
Restore table
Create snapshot
Configure cross-region snapshot
Relocate
Permissions
Manage IAM roles
Change admin user password
Manage tags
General information
Cluster performance
Query monitoring
S
rties

Cluster identifier
flight-2016

Status
Available

Cluster namespace
[redacted]

Date created
[redacted]

Cluster configuration
Production

Storage used
0.25% (0.41 of 160

Multi-AZ
No

Endpoint
[redacted]

JDBC URL
[redacted]

ODBC URL
Driver={Amazon Redshift (…

Note

By default, AWS Glue jobs pass Amazon Redshift temporary credentials that are created using the role that you specified to run the job. We do not recommend using these credentials. For security purposes, these credentials expire after 1 hour.

Set up the role for the AWS Glue job

The AWS Glue job needs a role to access the Amazon S3 bucket. You do not need IAM permissions for the Amazon Redshift cluster, your access is controlled by connectivity in Amazon VPC and your database credentials.

Set up Amazon VPC

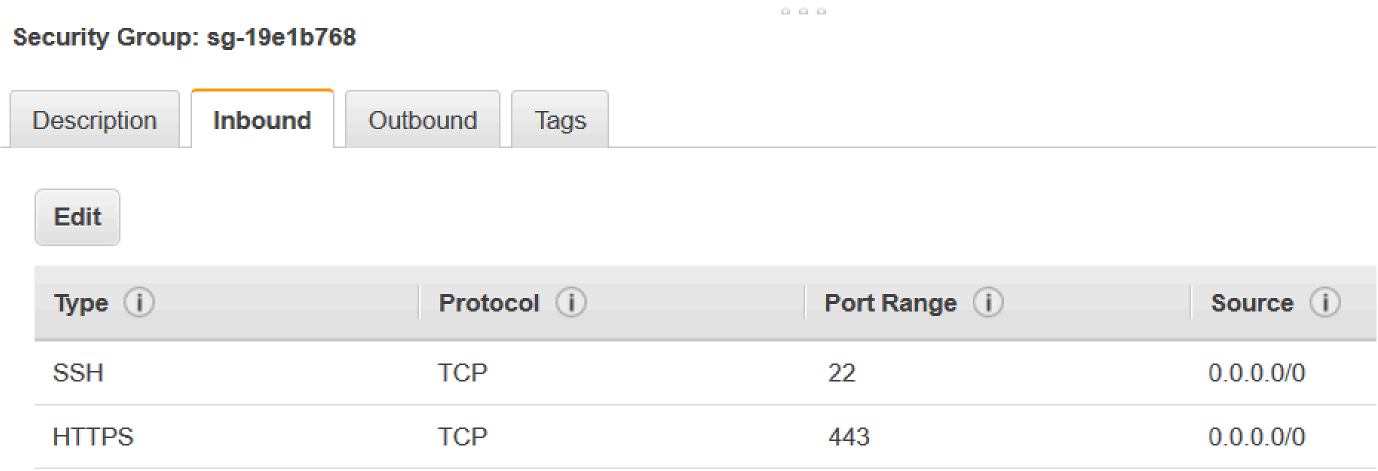
To set up access for Amazon Redshift data stores

1. Sign in to the AWS Management Console and open the Amazon Redshift console at https://console.aws.amazon.com/redshiftv2/.
2. In the left navigation pane, choose **Clusters**.
3. Choose the cluster name that you want to access from AWS Glue.
4. In the **Cluster Properties** section, choose a security group in **VPC security groups** to allow AWS Glue to use. Record the name of the security group that you chose for future reference. Choosing the security group opens the Amazon EC2 console **Security Groups** list.
5. Choose the security group to modify and navigate to the **Inbound** tab.
6. Add a self-referencing rule to allow AWS Glue components to communicate. Specifically, add or confirm that there is a rule of **Type** `All TCP`, **Protocol** is TCP, **Port Range** includes all ports, and whose **Source** is the same security group name as the **Group ID**.

   The inbound rule looks similar to the following:

Type	Protocol	Port range	Source
All TCP	TCP	0–65535	*database-security-group*

   For example:



Security Group: sg-19e1b768

Type	Protocol	Port Range	Source
SSH	TCP	22	0.0.0.0/0
HTTPS	TCP	443	0.0.0.0/0

7. Add a rule for outbound traffic also. Either open outbound traffic to all ports, for example:

Type	Protocol	Port range	Destination
All Traffic	ALL	ALL	0.0.0.0/0

Or create a self-referencing rule where **Type** `All TCP`, **Protocol** is TCP, **Port Range** includes all ports, and whose **Destination** is the same security group name as the **Group ID**. If using an Amazon S3 VPC endpoint, also add an HTTPS rule for Amazon S3 access. The *`s3-prefix-list-id`* is required in the security group rule to allow traffic from the VPC to the Amazon S3 VPC endpoint.

For example:

Type	Protocol	Port range	Destination
All TCP	TCP	0–65535	*`security-group`*
HTTPS	TCP	443	*`s3-prefix-list-id`*

Set up AWS Glue

You will need to create an AWS Glue Data Catalog connection that provides Amazon VPC connection information.

To configure Amazon Redshift Amazon VPC connectivity to AWS Glue in the console

1. Create a Data Catalog connection by following the steps in: the section called “Adding an AWS Glue connection”. After creating the connection, keep the connection name, *connectionName*, for the next step.

    - When selecting a **Connection type**, select **Amazon Redshift**.
    - When selecting a **Redshift cluster**, select your cluster by name.
    - Provide default connection information for a Amazon Redshift user on your cluster.
    - Your Amazon VPC settings will be automatically configured.

    > **Note**
    >
    > You will need to manually provide `PhysicalConnectionRequirements` for your Amazon VPC when creating an **Amazon Redshift** connection through the AWS SDK.

2. In your AWS Glue job configuration, provide *connectionName* as an **Additional network connection**.

Example: Reading from Amazon Redshift tables

You can read from Amazon Redshift clusters and Amazon Redshift serverless environments.

Prerequisites: An Amazon Redshift table you would like to read from. Follow the steps in the previous section the section called “Configure Redshift” after which you should have the Amazon S3 URI for a temporary directory, *temp-s3-dir* and an IAM role, *rs-role-name*, (in account *role-account-id*).

Using the Data Catalog

Additional Prerequisites: A Data Catalog Database and Table for the Amazon Redshift table you would like to read from. For more information about Data Catalog, see *Data discovery and cataloging*. After creating a entry for your Amazon Redshift table you will identify your connection with a *redshift-dc-database-name* and *redshift-table-name*.

Configuration: In your function options you will identify your Data Catalog Table with the `database` and `table_name` parameters. You will identify your Amazon S3 temporary directory with `redshift_tmp_dir`. You will also provide *`rs-role-name`* using the `aws_iam_role` key in the `additional_options` parameter.

```
glueContext.create_dynamic_frame.from_catalog(
    database = "redshift-dc-database-name",
    table_name = "redshift-table-name",
    redshift_tmp_dir = args["temp-s3-dir"],
    additional_options = {"aws_iam_role": "arn:aws:iam::role-account-id:role/rs-
role-name"})
```

Connecting directly

Additional Prerequisites: You will need the name of your Amazon Redshift table (*`redshift-table-name`*. You will need the JDBC connection information for the Amazon Redshift cluster storing that table. You will supply your connection information with *`host`*, *`port`*, *`redshift-database-name`*, *`username`* and *`password`*.

You can retrieve your connection information from the Amazon Redshift console when working with Amazon Redshift clusters. When using Amazon Redshift serverless, consult Connecting to Amazon Redshift Serverless in the Amazon Redshift documentation.

Configuration: In your function options you will identify your connection parameters with `url`, `dbtable`, `user` and `password`. You will identify your Amazon S3 temporary directory with `redshift_tmp_dir`. You can specify your IAM role using `aws_iam_role` when you use `from_options`. The syntax is similar to connecting through the Data Catalog, but you put the parameters in the `connection_options` map.

It is bad practice to hardcode passwords into AWS Glue scripts. Consider storing your passwords in AWS Secrets Manager and retrieving them in your script with SDK for Python (Boto3).

```
my_conn_options = {
    "url": "jdbc:redshift://host:port/redshift-database-name",
    "dbtable": "redshift-table-name",
    "user": "username",
    "password": "password",
    "redshiftTmpDir": args["temp-s3-dir"],
    "aws_iam_role": "arn:aws:iam::account id:role/rs-role-name"
}

df = glueContext.create_dynamic_frame.from_options("redshift", my_conn_options)
```

Example: Writing to Amazon Redshift tables

You can write to Amazon Redshift clusters and Amazon Redshift serverless environments.

Prerequisites: An Amazon Redshift cluster and follow the steps in the previous section the section called "Configure Redshift" after which you should have the Amazon S3 URI for a temporary directory, *temp-s3-dir* and an IAM role, *rs-role-name*, (in account *role-account-id*). You will also need a `DynamicFrame` whose contents you would like to write to the database.

Using the Data Catalog

Additional Prerequisites A Data Catalog Database for the Amazon Redshift cluster and table you would like to write to. For more information about Data Catalog, see *Data discovery and cataloging*. You will identify your connection with *redshift-dc-database-name* and the target table with *redshift-table-name*.

Configuration: In your function options you will identify your Data Catalog Database with the `database` parameter, then provide table with `table_name`. You will identify your Amazon S3 temporary directory with `redshift_tmp_dir`. You will also provide *rs-role-name* using the `aws_iam_role` key in the `additional_options` parameter.

```
glueContext.write_dynamic_frame.from_catalog(
    frame = input dynamic frame,
    database = "redshift-dc-database-name",
    table_name = "redshift-table-name",
    redshift_tmp_dir = args["temp-s3-dir"],
    additional_options = {"aws_iam_role": "arn:aws:iam::account-id:role/rs-role-
name"})
```

Connecting through a AWS Glue connection

You can connect to Amazon Redshift directly using the `write_dynamic_frame.from_options` method. However, rather than insert your connection details directly into your script, you can reference connection details stored in a Data Catalog connection with the `from_jdbc_conf` method. You can do this without crawling or creating Data Catalog tables for your database. For more information about Data Catalog connections, see *Connecting to data*.

Additional Prerequisites: A Data Catalog connection for your database, a Amazon Redshift table you would like to read from

Configuration: you will identify your Data Catalog connection with *`dc-connection-name`*. You will identify your Amazon Redshift database and table with *`redshift-table-name`* and *`redshift-database-name`*. You will provide your Data Catalog connection information with `catalog_connection` and your Amazon Redshift information with `dbtable` and `database`. The syntax is similar to connecting through the Data Catalog, but you put the parameters in the `connection_options` map.

```
my_conn_options = {
    "dbtable": "redshift-table-name",
    "database": "redshift-database-name",
    "aws_iam_role": "arn:aws:iam::role-account-id:role/rs-role-name"
}

glueContext.write_dynamic_frame.from_jdbc_conf(
    frame = input dynamic frame,
    catalog_connection = "dc-connection-name",
    connection_options = my_conn_options,
    redshift_tmp_dir = args["temp-s3-dir"])
```

Amazon Redshift connection option reference

The basic connection options used for all AWS Glue JDBC connections to set up information like `url`, `user` and `password` are consistent across all JDBC types. For more information about standard JDBC parameters, see the section called “JDBC connection parameters”.

The Amazon Redshift connection type takes some additional connection options:

- `"redshiftTmpDir"`: (Required) The Amazon S3 path where temporary data can be staged when copying out of the database.

- `"aws_iam_role"`: (Optional) ARN for an IAM role. The AWS Glue job will pass this role to the Amazon Redshift cluster to grant the cluster permissions needed to complete instructions from the job.

Additional connection options available in AWS Glue 4.0+

You can also pass options for the new Amazon Redshift connector through AWS Glue connection options. For a complete list of supported connector options, see the *Spark SQL parameters* section in Amazon Redshift integration for Apache Spark.

For you convenience, we reiterate certain new options here:

Name	Required	Default	Description
autopushdown	No	TRUE	Applies predicate and query pushdown by capturing and analyzing the Spark logical plans for SQL operations. The operations are translated into a SQL query, and then run in Amazon Redshift to improve performance.
autopushdown.s3_result_cache	No	FALSE	Caches the SQL query to unload data for Amazon S3 path mapping in memory so that the same query doesn't need to run again in the same Spark session. Only supported when

Name	Required	Default	Description
			`autopushdown` is enabled.
unload_s3_format	No	PARQUET	PARQUET - Unloads the query results in Parquet format. TEXT - Unloads the query results in pipe-delimited text format.
sse_kms_key	No	N/A	The AWS SSE-KMS key to use for encryption during `UNLOAD` operations instead of the default encryption for AWS.

Name	Required	Default	Description
extracopyoptions	No	N/A	A list of extra options to append to the Amazon Redshift `COPY`command when loading data, such as `TRUNCATECOLUMNS` or `MAXERROR n` (for other options see [COPY: Optional parameters](#)). Note that because these options are appended to the end of the `COPY` command, only options that make sense at the end of the command can be used. That should cover most possible use cases.
csvnullstring (experimental)	No	NULL	The String value to write for nulls when using the CSV `tempformat` . This should be a value that doesn't appear in your actual data.

These new parameters can be used in the following ways.

New options for performance improvement

The new connector introduces some new performance improvement options:

- `autopushdown`: Enabled by default.
- `autopushdown.s3_result_cache`: Disabled by default.
- `unload_s3_format`: `PARQUET` by default.

For information about using these options, see Amazon Redshift integration for Apache Spark. We recommend that you don't turn on `autopushdown.s3_result_cache` when you have mixed read and write operations because the cached results might contain stale information. The option `unload_s3_format` is set to `PARQUET` by default for the `UNLOAD` command, to improve performance and reduce storage cost. To use the `UNLOAD` command default behavior, reset the option to `TEXT`.

New encryption option for reading

By default, the data in the temporary folder that AWS Glue uses when it reads data from the Amazon Redshift table is encrypted using `SSE-S3` encryption. To use customer managed keys from AWS Key Management Service (AWS KMS) to encrypt your data, you can set up `("sse_kms_key" # kmsKey)` where ksmKey is the key ID from AWS KMS, instead of the legacy setting option `("extraunloadoptions" # s"ENCRYPTED KMS_KEY_ID '$kmsKey'")` in AWS Glue version 3.0.

```
datasource0 = glueContext.create_dynamic_frame.from_catalog(
  database = "database-name",
  table_name = "table-name",
  redshift_tmp_dir = args["TempDir"],
  additional_options = {"sse_kms_key":"<KMS_KEY_ID>"},
  transformation_ctx = "datasource0"
)
```

Support IAM-based JDBC URL

The new connector supports an IAM-based JDBC URL so you don't need to pass in a user/password or secret. With an IAM-based JDBC URL, the connector uses the job runtime role to access to the Amazon Redshift data source.

Step 1: Attach the following minimal required policy to your AWS Glue job runtime role.

JSON

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "VisualEditor0",
            "Effect": "Allow",
            "Action": "redshift:GetClusterCredentials",
            "Resource": [
                "arn:aws:redshift:us-east-1:111122223333:dbgroup:<cluster name>/
*",
                "arn:aws:redshift:*:111122223333:dbuser:*/*",
                "arn:aws:redshift:us-east-1:111122223333:dbname:<cluster name>/
<database name>"
            ]
        },
        {
            "Sid": "VisualEditor1",
            "Effect": "Allow",
            "Action": "redshift:DescribeClusters",
            "Resource": "*"
        }
    ]
}
```

Step 2: Use the IAM-based JDBC URL as follows. Specify a new option `DbUser` with the Amazon Redshift user name that you're connecting with.

```
conn_options = {
     // IAM-based JDBC URL
    "url": "jdbc:redshift:iam://<cluster name>:<region>/<database name>",
    "dbtable": dbtable,
    "redshiftTmpDir": redshiftTmpDir,
    "aws_iam_role": aws_iam_role,
    "DbUser": "<Redshift User name>" // required for IAM-based JDBC URL
    }

redshift_write = glueContext.write_dynamic_frame.from_options(
    frame=dyf,
    connection_type="redshift",
    connection_options=conn_options
)

redshift_read = glueContext.create_dynamic_frame.from_options(
    connection_type="redshift",
    connection_options=conn_options
)
```

> **Note**
>
> A `DynamicFrame` currently only supports an IAM-based JDBC URL with a `DbUser` in the `GlueContext.create_dynamic_frame.from_options` workflow.

Migrating from AWS Glue version 3.0 to version 4.0

In AWS Glue 4.0, ETL jobs have access to a new Amazon Redshift Spark connector and a new JDBC driver with different options and configuration. The new Amazon Redshift connector and driver are written with performance in mind, and keep transactional consistency of your data. These products are documented in the Amazon Redshift documentation. For more information, see:

- Amazon Redshift integration for Apache Spark
- Amazon Redshift JDBC driver, version 2.1

Table/column names and identifiers restriction

The new Amazon Redshift Spark connector and driver have a more restricted requirement for the Redshift table name. For more information, see Names and identifiers to define your Amazon Redshift table name. The job bookmark workflow might not work with a table name that doesn't match the rules and with certain characters, such as a space.

If you have legacy tables with names that don't conform to the Names and identifiers rules and see issues with bookmarks (jobs reprocessing old Amazon Redshift table data), we recommend that you rename your table names. For more information, see ALTER TABLE examples.

Default tempformat change in Dataframe

The AWS Glue version 3.0 Spark connector defaults the `tempformat` to CSV while writing to Amazon Redshift. To be consistent, in AWS Glue version 3.0, the `DynamicFrame` still defaults the `tempformat` to use `CSV`. If you've previously used Spark Dataframe APIs directly with the Amazon

Redshift Spark connector, you can explicitly set the `tempformat` to CSV in the `DataframeReader/Writer` options. Otherwise, `tempformat` defaults to `AVRO` in the new Spark connector.

Behavior change: map Amazon Redshift data type REAL to Spark data type FLOAT instead of DOUBLE

In AWS Glue version 3.0, Amazon Redshift `REAL` is converted to a Spark `DOUBLE` type. The new Amazon Redshift Spark connector has updated the behavior so that the Amazon Redshift `REAL` type is converted to, and back from, the Spark `FLOAT` type. If you have a legacy use case where you still want the Amazon Redshift `REAL` type to be mapped to a Spark `DOUBLE` type, you can use the following workaround:

- For a `DynamicFrame`, map the `Float` type to a `Double` type with `DynamicFrame.ApplyMapping`. For a `Dataframe`, you need to use `cast`.

Code example:

```
dyf_cast = dyf.apply_mapping([('a', 'long', 'a', 'long'), ('b', 'float', 'b',
 'double')])
```

Handling VARBYTE Data Type

When working with AWS Glue 3.0 and Amazon Redshift data types, AWS Glue 3.0 converts Amazon Redshift `VARBYTE` to Spark `STRING` type. However, the latest Amazon Redshift Spark connector doesn't support the `VARBYTE` data type. To work around this limitation, you can create a Redshift view that transforms `VARBYTE` columns to a supported data type. Then, use the new connector to load data from this view instead of the original table, which ensures compatibility while maintaining access to your `VARBYTE` data.

Example for Redshift query:

```
CREATE VIEW view_name AS SELECT FROM_VARBYTE(varbyte_column, 'hex') FROM table_name
```

Kafka connections

You can use a Kafka connection to read and write to Kafka data streams using information stored in a Data Catalog table, or by providing information to directly access the data stream. The connection supports a Kafka cluster or an Amazon Managed Streaming for Apache Kafka cluster. You can read information from Kafka into a Spark DataFrame, then convert it to a AWS Glue

DynamicFrame. You can write DynamicFrames to Kafka in a JSON format. If you directly access the data stream, use these options to provide the information about how to access the data stream.

If you use `getCatalogSource` or `create_data_frame_from_catalog` to consume records from a Kafka streaming source, or `getCatalogSink` or `write_dynamic_frame_from_catalog` to write records to Kafka, and the job has the Data Catalog database and table name information, and can use that to obtain some basic parameters for reading from the Kafka streaming source. If you use `getSource`, `getCatalogSink`, `getSourceWithFormat`, `getSinkWithFormat`, `createDataFrameFromOptions` or `create_data_frame_from_options`, or `write_dynamic_frame_from_catalog`, you must specify these basic parameters using the connection options described here.

You can specify the connection options for Kafka using the following arguments for the specified methods in the `GlueContext` class.

- Scala
  - `connectionOptions`: Use with `getSource`, `createDataFrameFromOptions`, `getSink`
  - `additionalOptions`: Use with `getCatalogSource`, `getCatalogSink`
  - `options`: Use with `getSourceWithFormat`, `getSinkWithFormat`
- Python
  - `connection_options`: Use with `create_data_frame_from_options`, `write_dynamic_frame_from_options`
  - `additional_options`: Use with `create_data_frame_from_catalog`, `write_dynamic_frame_from_catalog`
  - `options`: Use with `getSource`, `getSink`

For notes and restrictions about streaming ETL jobs, consult the section called "Streaming ETL notes and restrictions".

Configure Kafka

There are no AWS prerequisites to connecting to Kafka streams available through the internet.

You can create a AWS Glue Kafka connection to manage your connection credentials. For more information, see the section called "Creating a connection for a Kafka data stream". In your AWS Glue job configuration, provide *`connectionName`* as an **Additional network connection**, then, in your method call, provide *`connectionName`* to the `connectionName` parameter.

In certain cases, you will need to configure additional prerequisites:

- If using Amazon Managed Streaming for Apache Kafka with IAM authentication, you will need appropriate IAM configuration.
- If using Amazon Managed Streaming for Apache Kafka within an Amazon VPC, you will need appropriate Amazon VPC configuration. You will need to create a AWS Glue connection that provides Amazon VPC connection information. You will need your job configuration to include the AWS Glue connection as an **Additional network connection**.

For more information about Streaming ETL job prerequisites, consult the section called “Streaming ETL jobs”.

Example: Reading from Kafka streams

Used in conjunction with the section called “forEachBatch”.

Example for Kafka streaming source:

```
kafka_options =
    { "connectionName": "ConfluentKafka",
      "topicName": "kafka-auth-topic",
      "startingOffsets": "earliest",
      "inferSchema": "true",
      "classification": "json"
    }
data_frame_datasource0 =
 glueContext.create_data_frame.from_options(connection_type="kafka",
 connection_options=kafka_options)
```

Example: Writing to Kafka streams

Examples for writing to Kafka:

Example with the `getSink` method:

```
data_frame_datasource0 =
glueContext.getSink(
 connectionType="kafka",
 connectionOptions={
  JsonOptions("""{
   "connectionName": "ConfluentKafka",
   "classification": "json",
   "topic": "kafka-auth-topic",
   "typeOfData": "kafka"}
  """)},
 transformationContext="dataframe_ApacheKafka_node1711729173428")
 .getDataFrame()
```

Example with the `write_dynamic_frame.from_options` method:

```
kafka_options =
    { "connectionName": "ConfluentKafka",
      "topicName": "kafka-auth-topic",
      "classification": "json"
    }
data_frame_datasource0 =
 glueContext.write_dynamic_frame.from_options(connection_type="kafka",
 connection_options=kafka_options)
```

Kafka connection option reference

When reading, use the following connection options with `"connectionType": "kafka"`:

- `"bootstrap.servers"` (Required) A list of bootstrap server URLs, for example, as `b-1.vpc-test-2.o4q88o.c6.kafka.us-east-1.amazonaws.com:9094`. This option must be specified in the API call or defined in the table metadata in the Data Catalog.
- `"security.protocol"` (Required) The protocol used to communicate with brokers. The possible values are `"SSL"` or `"PLAINTEXT"`.
- `"topicName"` (Required) A comma-separated list of topics to subscribe to. You must specify one and only one of `"topicName"`, `"assign"` or `"subscribePattern"`.
- `"assign"`: (Required) A JSON string specifying the specific `TopicPartitions` to consume. You must specify one and only one of `"topicName"`, `"assign"` or `"subscribePattern"`.

  Example: '{"topicA":[0,1],"topicB":[2,4]}'
- `"subscribePattern"`: (Required) A Java regex string that identifies the topic list to subscribe to. You must specify one and only one of `"topicName"`, `"assign"` or `"subscribePattern"`.

  Example: 'topic.*'
- `"classification"` (Required) The file format used by the data in the record. Required unless provided through the Data Catalog.

- `"delimiter"` (Optional) The value separator used when `classification` is CSV. Default is "`,`."
- `"startingOffsets"`: (Optional) The starting position in the Kafka topic to read data from. The possible values are `"earliest"` or `"latest"`. The default value is `"latest"`.
- `"startingTimestamp"`: (Optional, supported only for AWS Glue version 4.0 or later) The Timestamp of the record in the Kafka topic to read data from. The possible value is a Timestamp string in UTC format in the pattern `yyyy-mm-ddTHH:MM:SSZ` (where `Z` represents a UTC timezone offset with a +/-. For example: "2023-04-04T08:00:00-04:00").

  Note: Only one of 'startingOffsets' or 'startingTimestamp' can be present in the Connection Options list of the AWS Glue streaming script, including both these properties will result in job failure.
- `"endingOffsets"`: (Optional) The end point when a batch query is ended. Possible values are either `"latest"` or a JSON string that specifies an ending offset for each `TopicPartition`.

  For the JSON string, the format is `{"topicA":{"0":23,"1":-1},"topicB":{"0":-1}}`. The value `-1` as an offset represents `"latest"`.
- `"pollTimeoutMs"`: (Optional) The timeout in milliseconds to poll data from Kafka in Spark job executors. The default value is `600000`.
- `"numRetries"`: (Optional) The number of times to retry before failing to fetch Kafka offsets. The default value is `3`.
- `"retryIntervalMs"`: (Optional) The time in milliseconds to wait before retrying to fetch Kafka offsets. The default value is `10`.
- `"maxOffsetsPerTrigger"`: (Optional) The rate limit on the maximum number of offsets that are processed per trigger interval. The specified total number of offsets is proportionally split across `topicPartitions` of different volumes. The default value is null, which means that the consumer reads all offsets until the known latest offset.
- `"minPartitions"`: (Optional) The desired minimum number of partitions to read from Kafka. The default value is null, which means that the number of spark partitions is equal to the number of Kafka partitions.
- `"includeHeaders"`: (Optional) Whether to include the Kafka headers. When the option is set to "true", the data output will contain an additional column named "glue_streaming_kafka_headers" with type `Array[Struct(key: String, value: String)]`. The default value is "false". This option is available in AWS Glue version 3.0 or later.

- `"schema"`: (Required when inferSchema set to false) The schema to use to process the payload. If classification is `avro` the provided schema must be in the Avro schema format. If the classification is not `avro` the provided schema must be in the DDL schema format.

  The following are schema examples.

  Example in DDL schema format

  ```
  'column1' INT, 'column2' STRING , 'column3' FLOAT
  ```

  Example in Avro schema format

  ```
  {
  "type":"array",
  "items":
  {
  "type":"record",
  "name":"test",
  "fields":
  [
    {
      "name":"_id",
      "type":"string"
    },
    {
      "name":"index",
      "type":
      [
        "int",
        "string",
        "float"
      ]
    }
  ]
  }
  }
  ```

- `"inferSchema"`: (Optional) The default value is 'false'. If set to 'true', the schema will be detected at runtime from the payload within `foreachbatch`.
- `"avroSchema"`: (Deprecated) Parameter used to specify a schema of Avro data when Avro format is used. This parameter is now deprecated. Use the `schema` parameter.

- `"addRecordTimestamp"`: (Optional) When this option is set to 'true', the data output will contain an additional column named "__src_timestamp" that indicates the time when the corresponding record received by the topic. The default value is 'false'. This option is supported in AWS Glue version 4.0 or later.
- `"emitConsumerLagMetrics"`: (Optional) When the option is set to 'true', for each batch, it will emit the metrics for the duration between the oldest record received by the topic and the time it arrives in AWS Glue to CloudWatch. The metric's name is "glue.driver.streaming.maxConsumerLagInMs". The default value is 'false'. This option is supported in AWS Glue version 4.0 or later.

When writing, use the following connection options with `"connectionType": "kafka"`:

- `"connectionName"` (Required) Name of the AWS Glue connection used to connect to the Kafka cluster (similar to Kafka source).
- `"topic"` (Required) If a topic column exists then its value is used as the topic when writing the given row to Kafka, unless the topic configuration option is set. That is, the `topic` configuration option overrides the topic column.
- `"partition"` (Optional) If a valid partition number is specified, that `partition` will be used when sending the record.

  If no partition is specified but a `key` is present, a partition will be chosen using a hash of the key.

  If neither `key` nor `partition` is present, a partition will be chosen based on sticky partitioning those changes when at least batch.size bytes are produced to the partition.
- `"key"` (Optional) Used for partitioning if `partition` is null.
- `"classification"` (Optional) The file format used by the data in the record. We only support JSON, CSV and Avro.

  With Avro format, we can provide a custom avroSchema to serialize with, but note that this needs to be provided on the source for deserializing as well. Else, by default it uses the Apache AvroSchema for serializing.

Additionally, you can fine-tune the Kafka sink as required by updating the Kafka producer configuration parameters. Note that there is no allow listing on connection options, all the key-value pairs are persisted on the sink as is.

However, there is a small deny list of options that will not take effect. For more information, see Kafka specific configurations.

Azure Cosmos DB connections

You can use AWS Glue for Spark to read from and write to existing containers in Azure Cosmos DB using the NoSQL API in AWS Glue 4.0 and later versions. You can define what to read from Azure Cosmos DB with a SQL query. You connect to Azure Cosmos DB using an Azure Cosmos DB Key stored in AWS Secrets Manager through a AWS Glue connection.

For more information about Azure Cosmos DB for NoSQL, consult the Azure documentation.

Configuring Azure Cosmos DB connections

To connect to Azure Cosmos DB from AWS Glue, you will need to create and store your Azure Cosmos DB Key in a AWS Secrets Manager secret, then associate that secret with a Azure Cosmos DB AWS Glue connection.

Prerequisites:

- In Azure, you will need to identify or generate an Azure Cosmos DB Key for use by AWS Glue, `cosmosKey`. For more information, see Secure access to data in Azure Cosmos DB in the Azure documentation.

To configure a connection to Azure Cosmos DB:

1. In AWS Secrets Manager, create a secret using your Azure Cosmos DB Key. To create a secret in Secrets Manager, follow the tutorial available in Create an AWS Secrets Manager secret in the AWS Secrets Manager documentation. After creating the secret, keep the Secret name, *`secretName`* for the next step.
    - When selecting **Key/value pairs**, create a pair for the key `spark.cosmos.accountKey` with the value *`cosmosKey`*.
2. In the AWS Glue console, create a connection by following the steps in the section called “Adding an AWS Glue connection”. After creating the connection, keep the connection name, *`connectionName`*, for future use in AWS Glue.
    - When selecting a **Connection type**, select Azure Cosmos DB.
    - When selecting an **AWS Secret**, provide *`secretName`*.

After creating a AWS Glue Azure Cosmos DB connection, you will need to perform the following steps before running your AWS Glue job:

- Grant the IAM role associated with your AWS Glue job permission to read *`secretName`*.
- In your AWS Glue job configuration, provide *`connectionName`* as an **Additional network connection**.

Reading from Azure Cosmos DB for NoSQL containers

Prerequisites:

- A Azure Cosmos DB for NoSQL container you would like to read from. You will need identification information for the container.

  An Azure Cosmos for NoSQL container is identified by its database and container. You must provide the database, *`cosmosDBName`*, and container, *`cosmosContainerName`*, names when connecting to the Azure Cosmos for NoSQL API.
- A AWS Glue Azure Cosmos DB connection configured to provide auth and network location information. To acquire this, complete the steps in the previous procedure, *To configure a connection to Azure Cosmos DB*. You will need the name of the AWS Glue connection, *`connectionName`*.

For example:

```
azurecosmos_read = glueContext.create_dynamic_frame.from_options(
    connection_type="azurecosmos",
    connection_options={
    "connectionName": connectionName,
    "spark.cosmos.database": cosmosDBName,
    "spark.cosmos.container": cosmosContainerName,
    }
)
```

You can also provide a SELECT SQL query, to filter the results returned to your DynamicFrame. You will need to configure `query`.

For example:

```
azurecosmos_read_query = glueContext.create_dynamic_frame.from_options(
    connection_type="azurecosmos",
    connection_options={
        "connectionName": "connectionName",
        "spark.cosmos.database": cosmosDBName,
        "spark.cosmos.container": cosmosContainerName,
        "spark.cosmos.read.customQuery": "query"
    }
)
```

Writing to Azure Cosmos DB for NoSQL containers

This example writes information from an existing DynamicFrame, `dynamicFrame` to Azure Cosmos DB. If the container already has information, AWS Glue will append data from your DynamicFrame. If the information in the container has a different schema from the information you write, you will run into errors.

Prerequisites:

- A Azure Cosmos DB table you would like to write to. You will need identification information for the container. **You must create the container before calling the connection method.**

  An Azure Cosmos for NoSQL container is identified by its database and container. You must provide the database, `cosmosDBName`, and container, `cosmosContainerName`, names when connecting to the Azure Cosmos for NoSQL API.
- A AWS Glue Azure Cosmos DB connection configured to provide auth and network location information. To acquire this, complete the steps in the previous procedure, *To configure a connection to Azure Cosmos DB*. You will need the name of the AWS Glue connection, `connectionName`.

For example:

```
azurecosmos_write = glueContext.write_dynamic_frame.from_options(
    frame=dynamicFrame,
    connection_type="azurecosmos",
    connection_options={
    "connectionName": connectionName,
    "spark.cosmos.database": cosmosDBName,
    "spark.cosmos.container": cosmosContainerName
)
```

Azure Cosmos DB connection option reference

- `connectionName` — Required. Used for Read/Write. The name of a AWS Glue Azure Cosmos DB connection configured to provide auth and network location information to your connection method.
- `spark.cosmos.database` — Required. Used for Read/Write. Valid Values: database names. Azure Cosmos DB for NoSQL database name.
- `spark.cosmos.container` — Required. Used for Read/Write. Valid Values: container names. Azure Cosmos DB for NoSQL container name.
- `spark.cosmos.read.customQuery` — Used for Read. Valid Values: SELECT SQL queries. Custom query to select documents to be read.

Azure SQL connections

You can use AWS Glue for Spark to read from and write to tables on Azure SQL Managed Instances in AWS Glue 4.0 and later versions. You can define what to read from Azure SQL with a SQL query. You connect to Azure SQL using user and password credentials stored in AWS Secrets Manager through a AWS Glue connection.

For more information about Azure SQL, consult the Azure SQL documentation.

Configuring Azure SQL connections

To connect to Azure SQL from AWS Glue, you will need to create and store your Azure SQL credentials in a AWS Secrets Manager secret, then associate that secret with a Azure SQL AWS Glue connection.

To configure a connection to Azure SQL:

1. In AWS Secrets Manager, create a secret using your Azure SQL credentials. To create a secret in Secrets Manager, follow the tutorial available in Create an AWS Secrets Manager secret in the AWS Secrets Manager documentation. After creating the secret, keep the Secret name, *`secretName`* for the next step.

   - When selecting **Key/value pairs**, create a pair for the key `user` with the value *`azuresqlUsername`*.
   - When selecting **Key/value pairs**, create a pair for the key `password` with the value *`azuresqlPassword`*.

2. In the AWS Glue console, create a connection by following the steps in the section called "Adding an AWS Glue connection". After creating the connection, keep the connection name, `connectionName`, for future use in AWS Glue.

   - When selecting a **Connection type**, select Azure SQL.
   - When providing **Azure SQL URL**, provide a JDBC endpoint URL.

     The URL must be in the following format: `jdbc:sqlserver://databaseServerName:databasePort;databaseName=azuresqlDBnam`

     AWS Glue requires the following URL properties:

     - `databaseName` – A default database in Azure SQL to connect to.

     For more information about JDBC URLs for Azure SQL Managed Instances, see the Microsoft documentation.
   - When selecting an **AWS Secret**, provide `secretName`.

After creating a AWS Glue Azure SQL connection, you will need to perform the following steps before running your AWS Glue job:

- Grant the IAM role associated with your AWS Glue job permission to read `secretName`.
- In your AWS Glue job configuration, provide `connectionName` as an **Additional network connection**.

Reading from Azure SQL tables

Prerequisites:

- A Azure SQL table you would like to read from. You will need identification information for the table, `databaseName` and `tableIdentifier`.

  An Azure SQL table is identified by its database, schema and table name. You must provide the database name and table name when connecting to Azure SQL. You also must provide the schema if it is not the default, "public". Database is provided through a URL property in `connectionName` , schema and table name through the `dbtable`.

- A AWS Glue Azure SQL connection configured to provide auth information. Complete the steps in the previous procedure, *To configure a connection to Azure SQL* to configure your auth information. You will need the name of the AWS Glue connection, *`connectionName`*.

For example:

```
azuresql_read_table = glueContext.create_dynamic_frame.from_options(
    connection_type="azuresql",
    connection_options={
        "connectionName": "connectionName",
        "dbtable": "tableIdentifier"
    }
)
```

You can also provide a SELECT SQL query, to filter the results returned to your DynamicFrame. You will need to configure `query`.

For example:

```
azuresql_read_query = glueContext.create_dynamic_frame.from_options(
    connection_type="azuresql",
    connection_options={
        "connectionName": "connectionName",
        "query": "query"
    }
)
```

Writing to Azure SQL tables

This example writes information from an existing DynamicFrame, *`dynamicFrame`* to Azure SQL. If the table already has information, AWS Glue will append data from your DynamicFrame.

Prerequisites:

- A Azure SQL table you would like to write to. You will need identification information for the table, *`databaseName`* and *`tableIdentifier`*.

  An Azure SQL table is identified by its database, schema and table name. You must provide the database name and table name when connecting to Azure SQL. You also must provide the schema if it is not the default, "public". Database is provided through a URL property in *`connectionName`* , schema and table name through the `dbtable`.

- Azure SQL auth information. Complete the steps in the previous procedure, *To configure a connection to Azure SQL* to configure your auth information. You will need the name of the AWS Glue connection, *`connectionName`*.

For example:

```
azuresql_write = glueContext.write_dynamic_frame.from_options(
    connection_type="azuresql",
    connection_options={
        "connectionName": "connectionName",
        "dbtable": "tableIdentifier"
    }
)
```

Azure SQL connection option reference

- `connectionName` — Required. Used for Read/Write. The name of a AWS Glue Azure SQL connection configured to provide auth information to your connection method.
- `databaseName` — Used for Read/Write. Valid Values: Azure SQL database names. The name of the database in Azure SQL to connect to.
- `dbtable` — Required for writing, required for reading unless `query` is provided. Used for Read/Write. Valid Values: Names of Azure SQL tables, or period separated schema/table name combinations. Used to specify the table and schema that identify the table to connect to. The default schema is "public". If your table is in a non-default schema, provide this information in the form *`schemaName`*`.`*`tableName`*.
- `query` — Used for Read. A Transact-SQL SELECT query defining what should be retrieved when reading from Azure SQL. For more information, see the Microsoft documentation.

BigQuery connections

You can use AWS Glue for Spark to read from and write to tables in Google BigQuery in AWS Glue 4.0 and later versions. You can read from BigQuery with a Google SQL query. You connect to BigQuery using credentials stored in AWS Secrets Manager through a AWS Glue connection.

For more information about Google BigQuery, see the Google Cloud BigQuery website.

Configuring BigQuery connections

To connect to Google BigQuery from AWS Glue, you will need to create and store your Google Cloud Platform credentials in a AWS Secrets Manager secret, then associate that secret with a Google BigQuery AWS Glue connection.

To configure a connection to BigQuery:

1. In Google Cloud Platform, create and identify relevant resources:
   - Create or identify a GCP project containing BigQuery tables you would like to connect to.
   - Enable the BigQuery API. For more information, see Use the BigQuery Storage Read API to read table data .
2. In Google Cloud Platform, create and export service account credentials:

   You can use the BigQuery credentials wizard to expedite this step: Create credentials.

   To create a service account in GCP, follow the tutorial available in Create service accounts.

   - When selecting **project**, select the project containing your BigQuery table.
   - When selecting GCP IAM roles for your service account, add or create a role that would grant appropriate permissions to run BigQuery jobs to read, write or create BigQuery tables.

   To create credentials for your service account, follow the tutorial available in Create a service account key.

   - When selecting key type, select **JSON**.

   You should now have downloaded a JSON file with credentials for your service account. It should look similar to the following:

```
{
  "type": "service_account",
  "project_id": "*****",
  "private_key_id": "*****",
  "private_key": "*****",
  "client_email": "*****",
  "client_id": "*****",
  "auth_uri": "https://accounts.google.com/o/oauth2/auth",
  "token_uri": "https://oauth2.googleapis.com/token",
  "auth_provider_x509_cert_url": "https://www.googleapis.com/oauth2/v1/certs",
  "client_x509_cert_url": "*****",
  "universe_domain": "googleapis.com"
}
```

3. base64 encode your downloaded credentials file. On an AWS CloudShell session or similar, you can do this from the command line by running `cat` *`credentialsFile.json`* `| base64 -w 0`. Retain the output of this command, *`credentialString`*.
4. In AWS Secrets Manager, create a secret using your Google Cloud Platform credentials. To create a secret in Secrets Manager, follow the tutorial available in Create an AWS Secrets Manager secret in the AWS Secrets Manager documentation. After creating the secret, keep the Secret name, *`secretName`* for the next step.

   - When selecting **Key/value pairs**, create a pair for the key `credentials` with the value *`credentialString`*.
5. In the AWS Glue Data Catalog, create a connection by following the steps in the section called “Adding an AWS Glue connection”. After creating the connection, keep the connection name, *`connectionName`*, for the next step.

   - When selecting a **Connection type**, select Google BigQuery.
   - When selecting an **AWS Secret**, provide *`secretName`*.
6. Grant the IAM role associated with your AWS Glue job permission to read *`secretName`*.
7. In your AWS Glue job configuration, provide *`connectionName`* as an **Additional network connection**.

Reading from BigQuery tables

Prerequisites:

- A BigQuery table you would like to read from. You will need the BigQuery table and dataset names, in the form `[dataset].[table]`. Let's call this *`tableName`*.
- The billing project for the BigQuery table. You will need the name of the project, *`parentProject`*. If there is no billing parent project, use the project containing the table.
- BigQuery auth information. Complete the steps *To manage your connection credentials with AWS Glue* to configure your auth information. You will need the name of the AWS Glue connection, *`connectionName`*.

For example:

```
bigquery_read = glueContext.create_dynamic_frame.from_options(
    connection_type="bigquery",
    connection_options={
        "connectionName": "connectionName",
        "parentProject": "parentProject",
        "sourceType": "table",
        "table": "tableName",
    }
```

You can also provide a query, to filter the results returned to your DynamicFrame. You will need to configure `query`, `sourceType`, `viewsEnabled` and `materializationDataset`.

For example:

Additional prerequisites:

You will need to create or identify a BigQuery dataset, *`materializationDataset`*, where BigQuery can write materialized views for your queries.

You will need to grant appropriate GCP IAM permissions to your service account to create tables in *`materializationDataset`*.

```
glueContext.create_dynamic_frame.from_options(
            connection_type="bigquery",
            connection_options={
                "connectionName": "connectionName",
                "materializationDataset": materializationDataset,
                "parentProject": "parentProject",
                "viewsEnabled": "true",
                "sourceType": "query",
                "query": "select * from bqtest.test"
            }
        )
```

Writing to BigQuery tables

This example writes directly to the BigQuery service. BigQuery also supports the "indirect" writing method. For more information about configuring indirect writes, see [the section called “Using indirect write with Google BigQuery”](#).

Prerequisites:

- A BigQuery table you would like to write to. You will need the BigQuery table and dataset names, in the form `[dataset].[table]`. You can also provide a new table name that will automatically be created. Let's call this *`tableName`*.
- The billing project for the BigQuery table. You will need the name of the project, *`parentProject`*. If there is no billing parent project, use the project containing the table.
- BigQuery auth information. Complete the steps *To manage your connection credentials with AWS Glue* to configure your auth information. You will need the name of the AWS Glue connection, *`connectionName`*.

For example:

```
bigquery_write = glueContext.write_dynamic_frame.from_options(
    frame=frameToWrite,
    connection_type="bigquery",
    connection_options={
        "connectionName": "connectionName",
        "parentProject": "parentProject",
        "writeMethod": "direct",
        "table": "tableName",
    }
)
```

BigQuery connection option reference

- `project` — Default: Google Cloud service account default. Used for Read/Write. The name of a Google Cloud project associated with your table.
- `table` — (Required) Used for Read/Write. The name of your BigQuery table in the format `[[project:]dataset.]`.
- `dataset` — Required when not defined through the `table` option. Used for Read/Write. The name of the dataset containing your BigQuery table.
- `parentProject` — Default: Google Cloud service account default. Used for Read/Write. The name of a Google Cloud project associated with `project` used for billing.
- `sourceType` — Used for Read. Required when reading. Valid Values: `table`, `query` Informs AWS Glue of whether you will read by table or by query.

- `materializationDataset` — Used for Read. Valid Values: strings. The name of a BigQuery dataset used to store materializations for views.
- `viewsEnabled` — Used for Read. Default: false. Valid Values: true, false. Configures whether BigQuery will use views.
- `query` — Used for Read. Used when `viewsEnabled` is true. A GoogleSQL DQL query.
- `temporaryGcsBucket` — Used for Write. Required when `writeMethod` is set to default (`indirect`). Name of a Google Cloud Storage bucket used to store an intermediate form of your data while writing to BigQuery.
- `writeMethod` — Default: `indirect`. Valid Values: `direct`, `indirect`. Used for Write. Specifies the method used to write your data.
  - If set to `direct`, your connector will write using the BigQuery Storage Write API.
  - If set to `indirect`, you connector will write to Google Cloud Storage, then transfer it to BigQuery using a load operation. Your Google Cloud service account will need appropriate GCS permissions.

Using indirect write with Google BigQuery

This example uses indirect write, which writes data to Google Cloud Storage and copies it to Google BigQuery.

Prerequisites:

You will need a temporary Google Cloud Storage bucket, *`temporaryBucket`*.

The GCP IAM role for AWS Glue's GCP service account will need appropriate GCS permissions to access *`temporaryBucket`*.

Additional Configuration:

To configure indirect write with BigQuery:

1. Assess the section called “Configuring BigQuery” and locate or redownload your GCP credentials JSON file. Identify *`secretName`*, the AWS Secrets Manager secret for the Google BigQuery AWS Glue connection used in your job.
2. Upload your credentials JSON file to an appropriately secure Amazon S3 location. Retain the path to the file, *`s3secretpath`* for future steps.

3. Edit *`secretName`*, adding the `spark.hadoop.google.cloud.auth.service.account.json.keyfile` key. Set the value to *`s3secretpath`*.

4. Grant your AWS Glue job Amazon S3 IAM permissions to access *`s3secretpath`*.

You can now provide your temporary GCS bucket location to your write method. You do not need to provide `writeMethod`, as `indirect` is historically the default.

```
bigquery_write = glueContext.write_dynamic_frame.from_options(
    frame=frameToWrite,
    connection_type="bigquery",
    connection_options={
        "connectionName": "connectionName",
        "parentProject": "parentProject",
        "temporaryGcsBucket": "temporaryBucket",
        "table": "tableName",
    }
)
```

JDBC connections

Certain, typically relational, database types support connecting through the JDBC standard. For more information about JDBC, see the Java JDBC API documentation. AWS Glue natively supports connecting to certain databases through their JDBC connectors - the JDBC libraries are provided in AWS Glue Spark jobs. When connecting to these database types using AWS Glue libraries, you have access to a standard set of options.

The JDBC connectionType values include the following:

- `"connectionType": "sqlserver"`: Designates a connection to a Microsoft SQL Server database.
- `"connectionType": "mysql"`: Designates a connection to a MySQL database.
- `"connectionType": "oracle"`: Designates a connection to an Oracle database.
- `"connectionType": "postgresql"`: Designates a connection to a PostgreSQL database.
- `"connectionType": "redshift"`: Designates a connection to an Amazon Redshift database. For more information, see the section called “Redshift connections”.

The following table lists the JDBC driver versions that AWS Glue supports.

Product	JDBC driver versions for Glue 5.0	JDBC driver versions for Glue 4.0	JDBC driver versions for Glue 3.0	JDBC driver versions for Glue 0.9, 1.0, 2.0
Microsoft SQL Server	10.2.0	9.4.0	7.x	6.x
MySQL	8.0.33	8.0.23	8.0.23	5.1
Oracle Database	23.3.0.23.09	21.7	21.1	11.2
PostgreSQL	42.7.3	42.3.6	42.2.18	42.1.x
Amazon Redshift *	redshift-jdbc42-2.1.0.29	redshift-jdbc42-2.1.0.16	redshift-jdbc41-1.2.12.1017	redshift-jdbc41-1.2.12.1017

* For the Amazon Redshift connection type, all other option name/value pairs that are included in connection options for a JDBC connection, including formatting options, are passed directly to the underlying SparkSQL DataSource. In AWS Glue with Spark jobs in AWS Glue 4.0 and later versions, the AWS Glue native connector for Amazon Redshift uses the Amazon Redshift integration for Apache Spark. For more information see Amazon Redshift integration for Apache Spark. In previous versions, see Amazon Redshift data source for Spark.

To configure your Amazon VPC to connect to Amazon RDS data stores using JDBC, refer to the section called "Setting up Amazon VPC to connect to Amazon RDS data stores".

Note

AWS Glue jobs are only associated with one subnet during a run. This may impact your ability to connect to multiple data sources through the same job. This behavior is not limited to JDBC sources.

Topics

- JDBC connection option reference
- Use sampleQuery
- Use custom JDBC driver
- Reading from JDBC tables in parallel
- Setting up Amazon VPC for JDBC connections to Amazon RDS data stores from AWS Glue

JDBC connection option reference

If you already have a JDBC AWS Glue connection defined, you can reuse the configuration properties defined in it, such as: url, user and password; so you don't have to specify them in the code as connection options. This feature is available in AWS Glue 3.0 and later versions. To do so, use the following connection properties:

- `"useConnectionProperties"`: Set it to "true" to indicate you want to use the configuration from a connection.
- `"connectionName"`: Enter the connection name to retrieve the configuration from, the connection must be defined in the same region as the job.

Use these connection options with JDBC connections:

- `"url"`: (Required) The JDBC URL for the database.
- `"dbtable"`: (Required) The database table to read from. For JDBC data stores that support schemas within a database, specify `schema.table-name`. If a schema is not provided, then the default "public" schema is used.
- `"user"`: (Required) The user name to use when connecting.
- `"password"`: (Required) The password to use when connecting.
- (Optional) The following options allow you to supply a custom JDBC driver. Use these options if you must use a driver that AWS Glue does not natively support.

  ETL jobs can use different JDBC driver versions for the data source and target, even if the source and target are the same database product. This allows you to migrate data between source and target databases with different versions. To use these options, you must first upload the JAR file of the JDBC driver to Amazon S3.

  - `"customJdbcDriverS3Path"`: The Amazon S3 path of the custom JDBC driver.
  - `"customJdbcDriverClassName"`: The class name of JDBC driver.

- `"bulkSize"`: (Optional) Used to configure parallel inserts for speeding up bulk loads into JDBC targets. Specify an integer value for the degree of parallelism to use when writing or inserting data. This option is helpful for improving the performance of writes into databases such as the Arch User Repository (AUR).
- `"hashfield"` (Optional) A string, used to specify the name of a column in the JDBC table to be used to divide the data into partitions when reading from JDBC tables in parallel. Provide "hashfield" OR "hashexpression". For more information, see the section called “Reading from JDBC in parallel”.
- `"hashexpression"` (Optional) A SQL select clause returning a whole number. Used to divide the data in a JDBC table into partitions when reading from JDBC tables in parallel. Provide "hashfield" OR "hashexpression". For more information, see the section called “Reading from JDBC in parallel”.
- `"hashpartitions"` (Optional) A positive integer. Used to specify the number of parallel reads of the JDBC table when reading from JDBC tables in parallel. Default: 7. For more information, see the section called “Reading from JDBC in parallel”.
- `"sampleQuery"`: (Optional) A custom SQL query statement. Used to specify a subset of information in a table to retrieve a sample of the table contents. **When configured without regard to your data, it can be less efficient than DynamicFrame methods, causing timeouts or out of memory errors.** For more information, see the section called “Use sampleQuery”.
- `"enablePartitioningForSampleQuery"`: (Optional) A boolean. Default: false. Used to enable reading from JDBC tables in parallel when specifying `sampleQuery`. **If set to true, `sampleQuery` must end with "where" or "and" for AWS Glue to append partitioning conditions.** For more information, see the section called “Use sampleQuery”.
- `"sampleSize"`: (Optional) A positive integer. Limits the number of rows returned by the sample query. Works only when `enablePartitioningForSampleQuery` is true. If partitioning is not enabled, you should instead directly add `"limit x"` in the `sampleQuery` to limit the size. For more information, see the section called “Use sampleQuery”.

Use sampleQuery

This section explains how to use `sampleQuery`, `sampleSize` and `enablePartitioningForSampleQuery`.

`sampleQuery` can be an efficient way to sample a few rows of your dataset. By default, the query is run by a single executor. When configured without regard to your data, it can be less efficient than DynamicFrame methods, causing timeouts or out of memory errors. Running SQL on the underlying database as part of your ETL pipeline is generally only needed for performance purposes. If you are trying to preview a few rows of your dataset, consider using the section called “show”. If you are trying to transform your dataset using SQL, consider using the section called “toDF” to define a SparkSQL transform against your data in a DataFrame form.

While your query may manipulate a variety of tables, `dbtable` remains required.

Using sampleQuery to retrieve a sample of your table

When using default sampleQuery behavior to retrieve a sample of your data, AWS Glue does not expect substantial throughput, so it runs your query on a single executor. In order to limit the data you provide and not cause performance problems, we suggest you provide SQL with a `LIMIT` clause.

Example Use sampleQuery without partitioning

The following code example shows how to use `sampleQuery` without partitioning.

```
//A full sql query statement.
val query = "select name from $tableName where age > 0 limit 1"
val connectionOptions = JsonOptions(Map(
    "url" -> url,
    "dbtable" -> tableName,
    "user" -> user,
    "password" -> password,
    "sampleQuery" -> query ))
val dyf = glueContext.getSource("mysql", connectionOptions)
          .getDynamicFrame()
```

Using sampleQuery against larger datasets

If you're reading a large dataset, you might need to enable JDBC partitioning to query a table in parallel. For more information, see the section called “Reading from JDBC in parallel”. To use `sampleQuery` with JDBC partitioning, set `enablePartitioningForSampleQuery` to true. Enabling this feature requires you to make some changes to your `sampleQuery`.

When using JDBC partitioning with `sampleQuery`, your query must end with "where" or "and" for AWS Glue to append partitioning conditions.

If you would like to limit the results of your sampleQuery when reading from JDBC tables in parallel, set the `"sampleSize"` parameter rather than specifying a `LIMIT` clause.

Example Use sampleQuery with JDBC partitioning

The following code example shows how to use `sampleQuery` with JDBC partitioning.

```
//note that the query should end with "where" or "and" if use with JDBC partitioning.
val query = "select name from $tableName where age > 0 and"

//Enable JDBC partitioning by setting hashfield.
//to use sampleQuery with partitioning, set enablePartitioningForSampleQuery.
//use sampleSize to limit the size of returned data.
val connectionOptions = JsonOptions(Map(
    "url" -> url,
    "dbtable" -> tableName,
    "user" -> user,
    "password" -> password,
    "hashfield" -> primaryKey,
    "sampleQuery" -> query,
    "enablePartitioningForSampleQuery" -> true,
    "sampleSize" -> "1" ))
val dyf = glueContext.getSource("mysql", connectionOptions)
          .getDynamicFrame()
```

Notes and Restrictions:

Sample queries cannot be used together with job bookmarks. The bookmark state will be ignored when configuration for both are provided.

Use custom JDBC driver

The following code examples show how to read from and write to JDBC databases with custom JDBC drivers. They demonstrate reading from one version of a database product, and writing to a later version of the same product.

Python

```
import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext, SparkConf
from awsglue.context import GlueContext
from awsglue.job import Job
import time
from pyspark.sql.types import StructType, StructField, IntegerType, StringType
```

```
sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session

# Construct JDBC connection options
connection_mysql5_options = {
    "url": "jdbc:mysql://<jdbc-host-name>:3306/db",
    "dbtable": "test",
    "user": "admin",
    "password": "pwd"}

connection_mysql8_options = {
    "url": "jdbc:mysql://<jdbc-host-name>:3306/db",
    "dbtable": "test",
    "user": "admin",
    "password": "pwd",
    "customJdbcDriverS3Path": "s3://amzn-s3-demo-bucket/mysql-connector-
java-8.0.17.jar",
    "customJdbcDriverClassName": "com.mysql.cj.jdbc.Driver"}

connection_oracle11_options = {
    "url": "jdbc:oracle:thin:@//<jdbc-host-name>:1521/ORCL",
    "dbtable": "test",
    "user": "admin",
    "password": "pwd"}

connection_oracle18_options = {
    "url": "jdbc:oracle:thin:@//<jdbc-host-name>:1521/ORCL",
    "dbtable": "test",
    "user": "admin",
    "password": "pwd",
    "customJdbcDriverS3Path": "s3://amzn-s3-demo-bucket/ojdbc10.jar",
    "customJdbcDriverClassName": "oracle.jdbc.OracleDriver"}

# Read from JDBC databases with custom driver
df_mysql8 = glueContext.create_dynamic_frame.from_options(connection_type="mysql",

 connection_options=connection_mysql8_options)

# Read DynamicFrame from MySQL 5 and write to MySQL 8
df_mysql5 = glueContext.create_dynamic_frame.from_options(connection_type="mysql",

 connection_options=connection_mysql5_options)
```

```
glueContext.write_from_options(frame_or_dfc=df_mysql5, connection_type="mysql",
                               connection_options=connection_mysql8_options)

# Read DynamicFrame from Oracle 11 and write to Oracle 18
df_oracle11 =
 glueContext.create_dynamic_frame.from_options(connection_type="oracle",

 connection_options=connection_oracle11_options)
glueContext.write_from_options(frame_or_dfc=df_oracle11, connection_type="oracle",
                               connection_options=connection_oracle18_options)
```

Scala

```
import com.amazonaws.services.glue.GlueContext
import com.amazonaws.services.glue.MappingSpec
import com.amazonaws.services.glue.errors.CallSite
import com.amazonaws.services.glue.util.GlueArgParser
import com.amazonaws.services.glue.util.Job
import com.amazonaws.services.glue.util.JsonOptions
import com.amazonaws.services.glue.DynamicFrame
import org.apache.spark.SparkContext
import scala.collection.JavaConverters._


object GlueApp {
  val MYSQL_5_URI: String = "jdbc:mysql://<jdbc-host-name>:3306/db"
  val MYSQL_8_URI: String = "jdbc:mysql://<jdbc-host-name>:3306/db"
  val ORACLE_11_URI: String = "jdbc:oracle:thin:@//<jdbc-host-name>:1521/ORCL"
  val ORACLE_18_URI: String = "jdbc:oracle:thin:@//<jdbc-host-name>:1521/ORCL"

  // Construct JDBC connection options
  lazy val mysql5JsonOption = jsonOptions(MYSQL_5_URI)
  lazy val mysql8JsonOption = customJDBCDriverJsonOptions(MYSQL_8_URI, "s3://amzn-
s3-demo-bucket/mysql-connector-java-8.0.17.jar", "com.mysql.cj.jdbc.Driver")
  lazy val oracle11JsonOption = jsonOptions(ORACLE_11_URI)
  lazy val oracle18JsonOption = customJDBCDriverJsonOptions(ORACLE_18_URI, "s3://
amzn-s3-demo-bucket/ojdbc10.jar", "oracle.jdbc.OracleDriver")

  def main(sysArgs: Array[String]): Unit = {
    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)
    val args = GlueArgParser.getResolvedOptions(sysArgs, Seq("JOB_NAME").toArray)
    Job.init(args("JOB_NAME"), glueContext, args.asJava)

    // Read from JDBC database with custom driver
    val df_mysql8: DynamicFrame = glueContext.getSource("mysql",
 mysql8JsonOption).getDynamicFrame()

    // Read DynamicFrame from MySQL 5 and write to MySQL 8
    val df_mysql5: DynamicFrame = glueContext.getSource("mysql",
 mysql5JsonOption).getDynamicFrame()
    glueContext.getSink("mysql", mysql8JsonOption).writeDynamicFrame(df_mysql5)

    // Read DynamicFrame from Oracle 11 and write to Oracle 18
    val df_oracle11: DynamicFrame = glueContext.getSource("oracle",
 oracle11JsonOption).getDynamicFrame()
    glueContext.getSink("oracle", oracle18JsonOption).writeDynamicFrame(df_oracle11)

    Job.commit()
  }

  private def jsonOptions(url: String): JsonOptions = {
    new JsonOptions(
      s"""{"url": "${url}",
         |"dbtable":"test",
         |"user": "admin",
         |"password": "pwd"}""".stripMargin)
  }

  private def customJDBCDriverJsonOptions(url: String, customJdbcDriverS3Path:
 String, customJdbcDriverClassName: String): JsonOptions = {
    new JsonOptions(
      s"""{"url": "${url}",
         |"dbtable":"test",
         |"user": "admin",
         |"password": "pwd",
         |"customJdbcDriverS3Path": "${customJdbcDriverS3Path}",
         |"customJdbcDriverClassName" :
 "${customJdbcDriverClassName}"}""".stripMargin)
  }
}
```

Reading from JDBC tables in parallel

You can set properties of your JDBC table to enable AWS Glue to read data in parallel. When you set certain properties, you instruct AWS Glue to run parallel SQL queries against logical partitions of your data. You can control partitioning by setting a hash field or a hash expression. You can also control the number of parallel reads that are used to access your data.

Reading from JDBC tables in parallel is an optimization technique that may improve performance. For more information about the process of identifying when this technique is appropriate, consult Reduce the amount of data scan in the *Best practices for performance tuning AWS Glue for Apache Spark jobs* guide on AWS Prescriptive Guidance.

To enable parallel reads, you can set key-value pairs in the parameters field of your table structure. Use JSON notation to set a value for the parameter field of your table. For more information about editing the properties of a table, see Viewing and managing table details. You can also enable parallel reads when you call the ETL (extract, transform, and load) methods `create_dynamic_frame_from_options` and `create_dynamic_frame_from_catalog`. For more information about specifying options in these methods, see from_options and from_catalog.

You can use this method for JDBC tables, that is, most tables whose base data is a JDBC data store. These properties are ignored when reading Amazon Redshift and Amazon S3 tables.

hashfield

Set `hashfield` to the name of a column in the JDBC table to be used to divide the data into partitions. For best results, this column should have an even distribution of values to spread the data between partitions. This column can be of any data type. AWS Glue generates non-overlapping queries that run in parallel to read the data partitioned by this column. For example, if your data is evenly distributed by month, you can use the `month` column to read each month of data in parallel.

```
'hashfield': 'month'
```

AWS Glue creates a query to hash the field value to a partition number and runs the query for all partitions in parallel. To use your own query to partition a table read, provide a `hashexpression` instead of a `hashfield`.

hashexpression

Set `hashexpression` to an SQL expression (conforming to the JDBC database engine grammar) that returns a whole number. A simple expression is the name of any numeric column in the table. AWS Glue generates SQL queries to read the JDBC data in parallel using the `hashexpression` in the `WHERE` clause to partition data.

For example, use the numeric column `customerID` to read data partitioned by a customer number.

```
'hashexpression': 'customerID'
```

To have AWS Glue control the partitioning, provide a `hashfield` instead of a `hashexpression`.

hashpartitions

Set `hashpartitions` to the number of parallel reads of the JDBC table. If this property is not set, the default value is 7.

For example, set the number of parallel reads to 5 so that AWS Glue reads your data with five queries (or fewer).

```
'hashpartitions': '5'
```

Setting up Amazon VPC for JDBC connections to Amazon RDS data stores from AWS Glue

When using JDBC to connect to databases in Amazon RDS, you will need to perform additional setup. To enable AWS Glue components to communicate with Amazon RDS, you must set up access to your Amazon RDS data stores in Amazon VPC. To enable AWS Glue to communicate between its components, specify a security group with a self-referencing inbound rule for all TCP ports. By creating a self-referencing rule, you can restrict the source to the same security group in the VPC. A self-referencing rule will not open the VPC to all networks. The default security group for your VPC might already have a self-referencing inbound rule for ALL Traffic.

To set up access between AWS Glue and Amazon RDS data stores

1. Sign in to the AWS Management Console and open the Amazon RDS console at https://console.aws.amazon.com/rds/.
2. In the Amazon RDS console, identify the security group(s) used to control access to your Amazon RDS database.

   In the left navigation pane, choose **Databases**, then select the instance you would like to connect to from the list in the main pane.

   In the database detail page, find **VPC security groups** on the **Connectivity & security** tab.
3. Based on your network architecture, identify which associated security group is best to modify to allow access for the AWS Glue service. Save its name, *`database-security-group`* for future reference. If there is no appropriate security group, follow the directions to Provide access to your DB instance in your VPC by creating a security group in the Amazon RDS documentation.
4. Sign in to the AWS Management Console and open the Amazon VPC console at https://console.aws.amazon.com/vpc/.
5. In the Amazon VPC console, identify how to update *`database-security-group`*.

   In the left navigation pane, choose **Security groups**, then select *`database-security-group`* from the list in the main pane.
6. Identify the security group ID for *`database-security-group`*, *`database-sg-id`*. Save it for future reference.

   In the security group detail page, find **Security group ID**.
7. Alter the inbound rules for *`database-security-group`*, add a self-referencing rule to allow AWS Glue components to communicate. Specifically, add or confirm that there is a rule where **Type** is `All TCP`, **Protocol** is `TCP`, **Port Range** includes all ports, and **Source** is *`database-sg-id`*. Verify that the security group you have entered for **Source** is the same as the security group you are editing.

   In the security group detail page, select **Edit inbound rules**.

   The inbound rule looks similar to this:

Type	Protocol	Port range	Source
All TCP	TCP	0–65535	*`database-sg-id`*

8. Add rules for outbound traffic.

   In the security group detail page, select **Edit outbound rules**.

   If you security group allows all outbound traffic, you do not need separate rules. For example:

Type	Protocol	Port range	Destination
All Traffic	ALL	ALL	0.0.0.0/0

   If your network architecture is designed for you to restrict outbound traffic, create the following outbound rules:

   Create a self-referencing rule where **Type** is `All TCP`, **Protocol** is `TCP`, **Port Range** includes all ports, and **Destination** is *`database-sg-id`*. Verify that the security group you have entered for **Destination** is the same as the security group you are editing.

   If using an Amazon S3 VPC endpoint, add an HTTPS rule to allow traffic from the VPC to Amazon S3. Create a rule where **Type** is `HTTPS`, **Protocol** is `TCP`, **Port Range** is `443` and **Destination** is the ID of the managed prefix list for the Amazon S3 gateway endpoint, *`s3-prefix-list-id`*. For more information about prefix lists and Amazon S3 gateway endpoints, see Gateway endpoints for Amazon S3 in the Amazon VPC documentation.

   For example:

Type	Protocol	Port range	Destination
All TCP	TCP	0–65535	*`database-sg-id`*
HTTPS	TCP	443	*`s3-prefix-list-id`*

MongoDB connections

You can use AWS Glue for Spark to read from and write to tables in MongoDB and MongoDB Atlas in AWS Glue 4.0 and later versions. You can connect to MongoDB using username and password credentials credentials stored in AWS Secrets Manager through a AWS Glue connection.

For more information about MongoDB, consult the MongoDB documentation.

Configuring MongoDB connections

To connect to MongoDB from AWS Glue, you will need your MongoDB credentials, *`mongodbUser`* and *`mongodbPass`*.

To connect to MongoDB from AWS Glue, you may need some prerequisites:

- If your MongoDB instance is in an Amazon VPC, configure Amazon VPC to allow your AWS Glue job to communicate with the MongoDB instance without traffic traversing the public internet.

  In Amazon VPC, identify or create a **VPC**, **Subnet** and **Security group** that AWS Glue will use while executing the job. Additionally, you need to ensure Amazon VPC is configured to permit network traffic between your MongoDB instance and this location. Based on your network layout, this may require changes to security group rules, Network ACLs, NAT Gateways and Peering connections.

You can then proceed to configure AWS Glue for use with MongoDB.

To configure a connection to MongoDB:

1. Optionally, in AWS Secrets Manager, create a secret using your MongoDB credentials. To create a secret in Secrets Manager, follow the tutorial available in Create an AWS Secrets Manager secret in the AWS Secrets Manager documentation. After creating the secret, keep the Secret name, *`secretName`* for the next step.

   - When selecting **Key/value pairs**, create a pair for the key `username` with the value *`mongodbUser`*.

     When selecting **Key/value pairs**, create a pair for the key `password` with the value *`mongodbPass`*.

2. In the AWS Glue console, create a connection by following the steps in the section called “Adding an AWS Glue connection”. After creating the connection, keep the connection name, *`connectionName`*, for future use in AWS Glue.

    - When selecting a **Connection type**, select **MongoDB** or **MongoDB Atlas**.
    - When selecting **MongoDB URL** or **MongoDB Atlas URL**, provide the hostname of your MongoDB instance.

      A MongoDB URL is provided in the format `mongodb://`*`mongoHost`*`:`*`mongoPort`*`/`*`mongoDBname`*.

      A MongoDB Atlas URL is provided in the format `mongodb+srv://`*`mongoHost`*`:`*`mongoPort`*`/`*`mongoDBname`*.

      Providing the default database for the connection, *`mongoDBname`* is optional.
    - If you chose to create an Secrets Manager secret, choose the AWS Secrets Manager **Credential type**.

      Then, in **AWS Secret** provide *`secretName`*.
    - If you choose to provide **Username and password**, provide *`mongodbUser`* and *`mongodbPass`*.
3. In the following situations, you may require additional configuration:

    - For MongoDB instances hosted on AWS in an Amazon VPC
        - You will need to provide Amazon VPC connection information to the AWS Glue connection that defines your MongoDB security credentials. When creating or updating your connection, set **VPC**, **Subnet** and **Security groups** in **Network options**.

After creating a AWS Glue MongoDB connection, you will need to perform the following actions before calling your connection method:

- If you chose to create an Secrets Manager secret, grant the IAM role associated with your AWS Glue job permission to read *`secretName`*.
- In your AWS Glue job configuration, provide *`connectionName`* as an **Additional network connection**.

To use your AWS Glue MongoDB connection in AWS Glue for Spark, provide the `connectionName` option in your connection method call. Alternatively, you can follow the steps in the section called "Integrating with MongoDB" to use the connection in conjunction with the AWS Glue Data Catalog.

Reading from MongoDB using a AWS Glue connection

Prerequisites:

- A MongoDB collection you would like to read from. You will need identification information for the collection.

  A MongoDB collection is identified by a database name and a collection name, *`mongodbName`*, *`mongodbCollection`*.
- A AWS Glue MongoDB connection configured to provide auth information. Complete the steps in the previous procedure, *To configure a connection to MongoDB* to configure your auth information. You will need the name of the AWS Glue connection, *`connectionName`*.

For example:

```
mongodb_read = glueContext.create_dynamic_frame.from_options(
    connection_type="mongodb",
    connection_options={
        "connectionName": "connectionName",
        "database": "mongodbName",
        "collection": "mongodbCollection",
        "partitioner":
 "com.mongodb.spark.sql.connector.read.partitioner.SinglePartitionPartitioner",
        "partitionerOptions.partitionSizeMB": "10",
        "partitionerOptions.partitionKey": "_id",
        "disableUpdateUri": "false",
    }
)
```

Writing to MongoDB tables

This example writes information from an existing DynamicFrame, *`dynamicFrame`* to MongoDB.

Prerequisites:

- A MongoDB collection you would like to write to. You will need identification information for the collection.

A MongoDB collection is identified by a database name and a collection name, *mongodbName*, *mongodbCollection*.

- A AWS Glue MongoDB connection configured to provide auth information. Complete the steps in the previous procedure, *To configure a connection to MongoDB* to configure your auth information. You will need the name of the AWS Glue connection, *connectionName*.

For example:

```
glueContext.write_dynamic_frame.from_options(
    frame=dynamicFrame,
    connection_type="mongodb",
    connection_options={
        "connectionName": "connectionName",
        "database": "mongodbName",
        "collection": "mongodbCollection",
        "disableUpdateUri": "false",
        "retryWrites": "false",
    },
)
```

Reading and writing to MongoDB tables

This example writes information from an existing DynamicFrame, *dynamicFrame* to MongoDB.

Prerequisites:

- A MongoDB collection you would like to read from. You will need identification information for the collection.

  A MongoDB collection you would like to write to. You will need identification information for the collection.

  A MongoDB collection is identified by a database name and a collection name, *mongodbName*, *mongodbCollection*.
- MongoDB auth information, *mongodbUser* and *mongodbPassword*.

For example:

Python

```
import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext, SparkConf
from awsglue.context import GlueContext
from awsglue.job import Job
import time

## @params: [JOB_NAME]
args = getResolvedOptions(sys.argv, ['JOB_NAME'])

sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session

job = Job(glueContext)
job.init(args['JOB_NAME'], args)

output_path = "s3://some_bucket/output/" + str(time.time()) + "/"
mongo_uri = "mongodb://<mongo-instanced-ip-address>:27017"
mongo_ssl_uri = "mongodb://<mongo-instanced-ip-address>:27017"
write_uri = "mongodb://<mongo-instanced-ip-address>:27017"

read_mongo_options = {
    "uri": mongo_uri,
    "database": "mongodbName",
    "collection": "mongodbCollection",
    "username": "mongodbUsername",
    "password": "mongodbPassword",
    "partitioner": "MongoSamplePartitioner",
    "partitionerOptions.partitionSizeMB": "10",
    "partitionerOptions.partitionKey": "_id"}

ssl_mongo_options = {
    "uri": mongo_ssl_uri,
    "database": "mongodbName",
    "collection": "mongodbCollection",
    "ssl": "true",
    "ssl.domain_match": "false"
}

write_mongo_options = {
    "uri": write_uri,
    "database": "mongodbName",
    "collection": "mongodbCollection",
    "username": "mongodbUsername",
    "password": "mongodbPassword",
}

# Get DynamicFrame from MongoDB
dynamic_frame =
 glueContext.create_dynamic_frame.from_options(connection_type="mongodb",

 connection_options=read_mongo_options)

# Write DynamicFrame to MongoDB
glueContext.write_dynamic_frame.from_options(dynamicFrame,
 connection_type="mongodb", connection_options=write_mongo_options)

job.commit()
```

Scala

```
import com.amazonaws.services.glue.GlueContext
import com.amazonaws.services.glue.MappingSpec
import com.amazonaws.services.glue.errors.CallSite
import com.amazonaws.services.glue.util.GlueArgParser
import com.amazonaws.services.glue.util.Job
import com.amazonaws.services.glue.util.JsonOptions
import com.amazonaws.services.glue.DynamicFrame
import org.apache.spark.SparkContext
import scala.collection.JavaConverters._

object GlueApp {
  val DEFAULT_URI: String = "mongodb://<mongo-instanced-ip-address>:27017"
  val WRITE_URI: String = "mongodb://<mongo-instanced-ip-address>:27017"
  lazy val defaultJsonOption = jsonOptions(DEFAULT_URI)
  lazy val writeJsonOption = jsonOptions(WRITE_URI)
  def main(sysArgs: Array[String]): Unit = {
    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)
    val args = GlueArgParser.getResolvedOptions(sysArgs, Seq("JOB_NAME").toArray)
    Job.init(args("JOB_NAME"), glueContext, args.asJava)

    // Get DynamicFrame from MongoDB
    val dynamicFrame: DynamicFrame = glueContext.getSource("mongodb",
 defaultJsonOption).getDynamicFrame()

    // Write DynamicFrame to MongoDB
    glueContext.getSink("mongodb", writeJsonOption).writeDynamicFrame(dynamicFrame)

    Job.commit()
  }

  private def jsonOptions(uri: String): JsonOptions = {
    new JsonOptions(
      s"""{"uri": "${uri}",
         |"database":"mongodbName",
         |"collection":"mongodbCollection",
         |"username": "mongodbUsername",
         |"password": "mongodbPassword",
         |"ssl":"true",
         |"ssl.domain_match":"false",
         |"partitioner": "MongoSamplePartitioner",
         |"partitionerOptions.partitionSizeMB": "10",
         |"partitionerOptions.partitionKey": "_id"}""".stripMargin)
  }
}
```

MongoDB connection option reference

Designates a connection to MongoDB. Connection options differ for a source connection and a sink connection.

These connection properties are shared between source and sink connections:

- `connectionName` — Used for Read/Write. The name of a AWS Glue MongoDB connection configured to provide auth and networking information to your connection method. When a AWS Glue connection is configured as described in the previous section, the section called "Configuring MongoDB", providing `connectionName` will replace the need to provide the `"uri"`, `"username"` and `"password"` connection options.
- `"uri"`: (Required) The MongoDB host to read from, formatted as `mongodb://<host>:<port>`. Used in AWS Glue versions prior to AWS Glue 4.0.
- `"connection.uri"`: (Required) The MongoDB host to read from, formatted as `mongodb://<host>:<port>`. Used in AWS Glue 4.0 and later versions.

- `"username"`: (Required) The MongoDB user name.
- `"password"`: (Required) The MongoDB password.
- `"database"`: (Required) The MongoDB database to read from. This option can also be passed in `additional_options` when calling `glue_context.create_dynamic_frame_from_catalog` in your job script.
- `"collection"`: (Required) The MongoDB collection to read from. This option can also be passed in `additional_options` when calling `glue_context.create_dynamic_frame_from_catalog` in your job script.

"connectionType": "mongodb" as source

Use the following connection options with `"connectionType": "mongodb"` as a source:

- `"ssl"`: (Optional) If `true`, initiates an SSL connection. The default is `false`.
- `"ssl.domain_match"`: (Optional) If `true` and `ssl` is `true`, domain match check is performed. The default is `true`.
- `"batchSize"`: (Optional): The number of documents to return per batch, used within the cursor of internal batches.
- `"partitioner"`: (Optional): The class name of the partitioner for reading input data from MongoDB. The connector provides the following partitioners:
  - `MongoDefaultPartitioner` (default) (Not supported in AWS Glue 4.0)
  - `MongoSamplePartitioner` (Requires MongoDB 3.2 or later) (Not supported in AWS Glue 4.0)
  - `MongoShardedPartitioner` (Not supported in AWS Glue 4.0)
  - `MongoSplitVectorPartitioner` (Not supported in AWS Glue 4.0)
  - `MongoPaginateByCountPartitioner` (Not supported in AWS Glue 4.0)
  - `MongoPaginateBySizePartitioner` (Not supported in AWS Glue 4.0)
  - `com.mongodb.spark.sql.connector.read.partitioner.SinglePartitionPartitioner`
  - `com.mongodb.spark.sql.connector.read.partitioner.ShardedPartitioner`
  - `com.mongodb.spark.sql.connector.read.partitioner.PaginateIntoPartitionsPartit`
- `"partitionerOptions"` (Optional): Options for the designated partitioner. The following options are supported for each partitioner:
  - `MongoSamplePartitioner`: `partitionKey`, `partitionSizeMB`, `samplesPerPartition`
  - `MongoShardedPartitioner`: `shardkey`

- `MongoSplitVectorPartitioner`: `partitionKey`, `partitionSizeMB`
- `MongoPaginateByCountPartitioner`: `partitionKey`, `numberOfPartitions`
- `MongoPaginateBySizePartitioner`: `partitionKey`, `partitionSizeMB`

  For more information about these options, see Partitioner Configuration in the MongoDB documentation.

"connectionType": "mongodb" as sink

Use the following connection options with `"connectionType": "mongodb"` as a sink:

- `"ssl"`: (Optional) If `true`, initiates an SSL connection. The default is `false`.
- `"ssl.domain_match"`: (Optional) If `true` and `ssl` is `true`, domain match check is performed. The default is `true`.
- `"extendedBsonTypes"`: (Optional) If `true`, allows extended BSON types when writing data to MongoDB. The default is `true`.
- `"replaceDocument"`: (Optional) If `true`, replaces the whole document when saving datasets that contain an `_id` field. If `false`, only fields in the document that match the fields in the dataset are updated. The default is `true`.
- `"maxBatchSize"`: (Optional): The maximum batch size for bulk operations when saving data. The default is 512.
- `"retryWrites"`: (Optional): Automatically retry certain write operations a single time if AWS Glue encounters a network error.

SAP HANA connections

You can use AWS Glue for Spark to read from and write to tables in SAP HANA in AWS Glue 4.0 and later versions. You can define what to read from SAP HANA with a SQL query. You connect to SAP HANA using JDBC credentials stored in AWS Secrets Manager through a AWS Glue SAP HANA connection.

For more information about SAP HANA JDBC, consult the SAP HANA documentation.

Configuring SAP HANA connections

To connect to SAP HANA from AWS Glue, you will need to create and store your SAP HANA credentials in a AWS Secrets Manager secret, then associate that secret with a SAP HANA AWS Glue connection. You will need to configure network connectivity between your SAP HANA service and AWS Glue.

To connect to SAP HANA, you may need some prerequisites:

- If your SAP HANA service is in an Amazon VPC, configure Amazon VPC to allow your AWS Glue job to communicate with the SAP HANA service without traffic traversing the public internet.

  In Amazon VPC, identify or create a **VPC**, **Subnet** and **Security group** that AWS Glue will use while executing the job. Additionally, you need to ensure Amazon VPC is configured to permit network traffic between your SAP HANA endpoint and this location. Your job will need to establish a TCP connection with your SAP HANA JDBC port. For more information about SAP HANA ports, see the SAP HANA documentation. Based on your network layout, this may require changes to security group rules, Network ACLs, NAT Gateways and Peering connections.
- There are no additional prerequisites if your SAP HANA endpoint is internet accesible.

To configure a connection to SAP HANA:

1. In AWS Secrets Manager, create a secret using your SAP HANA credentials. To create a secret in Secrets Manager, follow the tutorial available in Create an AWS Secrets Manager secret in the AWS Secrets Manager documentation. After creating the secret, keep the Secret name, *`secretName`* for the next step.
   - When selecting **Key/value pairs**, create a pair for the key `user` with the value *`saphanaUsername`*.
   - When selecting **Key/value pairs**, create a pair for the key `password` with the value *`saphanaPassword`*.
2. In the AWS Glue console, create a connection by following the steps in the section called “Adding an AWS Glue connection”. After creating the connection, keep the connection name, *`connectionName`*, for future use in AWS Glue.
   - When selecting a **Connection type**, select SAP HANA.
   - When providing **SAP HANA URL**, provide the URL for your instance.

     SAP HANA JDBC URLs are in the form `jdbc:sap://saphanaHostname:saphanaPort/?databaseName=saphanaDBname,Paramete`

     AWS Glue requires the following JDBC URL parameters:

  - `databaseName` – A default database in SAP HANA to connect to.
- When selecting an **AWS Secret**, provide *`secretName`*.

After creating a AWS Glue SAP HANA connection, you will need to perform the following steps before running your AWS Glue job:

- Grant the IAM role associated with your AWS Glue job permission to read *`secretName`*.
- In your AWS Glue job configuration, provide *`connectionName`* as an **Additional network connection**.

Reading from SAP HANA tables

Prerequisites:

- A SAP HANA table you would like to read from. You will need identification information for the table.

  A table can be specified with a SAP HANA table name and schema name, in the form *`schemaName`*`.`*`tableName`*. The schema name and "." separator are not required if the table is in the default schema, "public". Call this *`tableIdentifier`*. Note that the database is provided as a JDBC URL parameter in `connectionName`.
- A AWS Glue SAP HANA connection configured to provide auth information. Complete the steps in the previous procedure, *To configure a connection to SAP HANA* to configure your auth information. You will need the name of the AWS Glue connection, *`connectionName`*.

For example:

```
saphana_read_table = glueContext.create_dynamic_frame.from_options(
    connection_type="saphana",
    connection_options={
        "connectionName": "connectionName",
        "dbtable": "tableIdentifier",
    }
)
```

You can also provide a SELECT SQL query, to filter the results returned to your DynamicFrame. You will need to configure `query`.

For example:

```
saphana_read_query = glueContext.create_dynamic_frame.from_options(
    connection_type="saphana",
    connection_options={
        "connectionName": "connectionName",
        "query": "query"
    }
)
```

Writing to SAP HANA tables

This example writes information from an existing DynamicFrame, *`dynamicFrame`* to SAP HANA. If the table already has information, AWS Glue will error.

Prerequisites:

- A SAP HANA table you would like to write to.

  A table can be specified with a SAP HANA table name and schema name, in the form *`schemaName`*`.`*`tableName`*. The schema name and "." separator are not required if the table is in the default schema, "public". Call this *`tableIdentifier`*. Note that the database is provided as a JDBC URL parameter in `connectionName`.
- SAP HANA auth information. Complete the steps in the previous procedure, *To configure a connection to SAP HANA* to configure your auth information. You will need the name of the AWS Glue connection, *`connectionName`*.

For example:

```
options = {
    "connectionName": "connectionName",
    "dbtable": 'tableIdentifier'
}

    saphana_write = glueContext.write_dynamic_frame.from_options(
        frame=dynamicFrame,
        connection_type="saphana",
        connection_options=options
)
```

SAP HANA connection option reference

- `connectionName` — Required. Used for Read/Write. The name of a AWS Glue SAP HANA connection configured to provide auth and networking information to your connection method.
- `databaseName` — Used for Read/Write. Valid Values: names of databases in SAP HANA. Name of database to connect to.
- `dbtable` — Required for writing, required for reading unless `query` is provided. Used for Read/Write. Valid Values: contents of a SAP HANA SQL FROM clause. Identifies a table in SAP HANA to connect to. You may also provide other SQL than a table name, such as a subquery. For more information, see the From clause in the SAP HANA documentation.
- `query` — Used for Read. A SAP HANA SQL SELECT query defining what should be retrieved when reading from SAP HANA.

Snowflake connections

You can use AWS Glue for Spark to read from and write to tables in Snowflake in AWS Glue 4.0 and later versions. You can read from Snowflake with a SQL query. You can connect to Snowflake using a user and password. You can refer to Snowflake credentials stored in AWS Secrets Manager through the AWS Glue Data Catalog. Data Catalog Snowflake credentials for AWS Glue for Spark are stored separately from Data Catalog Snowflake credentials for crawlers. You must choose a `SNOWFLAKE` type connection and not a `JDBC` type connection configured to connect to Snowflake.

For more information about Snowflake, see the Snowflake website. For more information about Snowflake on AWS, see Snowflake Data Warehouse on Amazon Web Services.

Configuring Snowflake connections

There are no AWS prerequisites to connecting to Snowflake databases available through the internet.

Optionally, you can perform the following configuration to manage your connection credentials with AWS Glue.

To manage your connection credentials with AWS Glue

1. In Snowflake, generate a user, *`snowflakeUser`* and password, *`snowflakePassword`*.
2. In AWS Secrets Manager, create a secret using your Snowflake credentials. To create a secret in Secrets Manager, follow the tutorial available in Create an AWS Secrets Manager secret in

the AWS Secrets Manager documentation. After creating the secret, keep the Secret name, *`secretName`* for the next step.

- When selecting **Key/value pairs**, create a pair for *`snowflakeUser`* with the key `USERNAME`.
- When selecting **Key/value pairs**, create a pair for *`snowflakePassword`* with the key `PASSWORD`.
- When selecting **Key/value pairs**, you can provide your Snowflake warehouse with the key `sfWarehouse`.
- When selecting **Key/value pairs**, you can provide additional Snowflake connection properties using their corresponding Spark property names as keys. Supported properties include:
  - `sfDatabase` - Snowflake database name
  - `sfSchema` - Snowflake schema name
  - `sfRole` - Snowflake role name
  - `pem_private_key` - Private key for key-pair authentication

3. In the AWS Glue Data Catalog, create a connection by choosing **Connections**, then **Create connection**. Following the steps in the connection wizard to complete the process:

   - When selecting a **Data source**, select Snowflake, then choose **Next**.
   - Enter the connection details such as host and port. When entering the host **Snowflake URL**, provide the URL of your Snowflake instance. The URL will typically use a hostname in the form *`account_identifier`*`.snowflakecomputing.com`. However, the URL format may vary depending on your Snowflake account type (for example, AWS, Azure, or Snowflake-hosted).
   - When selecting the IAM service role, choose from the drop-down menu. This is the IAM role from your account that will be used to access AWS Secrets Manager and assign IP if VPC is specified.
   - When selecting an **AWS Secret**, provide *`secretName`*.

4. In the next step in the wizard, set properties for your Snowflake connection.
5. In the final step in the wizard, review your settings and then complete the process to create your connection.

In the following situations, you may require the following:

- For Snowflake hosted on AWS in an Amazon VPC
  - You will need appropriate Amazon VPC configuration for Snowflake. For more information on how to configure your Amazon VPC, consult AWS PrivateLink & Snowflake in the Snowflake documentation.
  - You will need appropriate Amazon VPC configuration for AWS Glue. the section called “Configuring interface VPC endpoints (AWS PrivateLink) for AWS Glue ”.
  - You will need to create a AWS Glue Data Catalog connection that provides Amazon VPC connection information (in addition to the id of an AWS Secrets Manager secret that defines your Snowflake security credentials). Your URL will change when using AWS PrivateLink, as described in the Snowflake documentation linked in a previous item.
  - You will need your job configuration in include the Data Catalog connection as an **Additional network connection**.

Reading from Snowflake tables

Prerequisites: A Snowflake table you would like to read from. You will need the Snowflake table name, *`tableName`*. You will need your Snowflake url *`snowflakeUrl`*, username *`snowflakeUser`* and password *`snowflakePassword`*. If your Snowflake user does not have a default namespace set, you will need the Snowflake database name, *`databaseName`* and the schema name *`schemaName`*. Additionally, if your Snowflake user does not have a default warehouse set, you will need a warehouse name *`warehouseName`*.

For example:

Additional Prerequisites: Complete the steps *To manage your connection credentials with AWS Glue* to configure *`snowflakeUrl`*, *`snowflakeUsername`* and *`snowflakePassword`*. To review these steps, see the section called “Configuring Snowflake”, the previous section. To select which **Additional network connection** to connect with, we will use the `connectionName` parameter.

```
snowflake_read = glueContext.create_dynamic_frame.from_options(
  connection_type="snowflake",
  connection_options={
        "connectionName": "connectionName",
        "dbtable": "tableName",
        "sfDatabase": "databaseName",
        "sfSchema": "schemaName",
        "sfWarehouse": "warehouseName",
    }
)
```

Additionally, you can use the `autopushdown` and `query` parameters to read a portion of a Snowflake table. This can be substantially more efficient than filtering your results after they have been loaded into Spark. Consider an example where all sales are stored in the same table, but you only need to analyze sales from a certain store on holidays. If that information is stored in the table, you could use predicate pushdown to retrieve the results as follows:

```
snowflake_node = glueContext.create_dynamic_frame.from_options(
    connection_type="snowflake",
    connection_options={
        "autopushdown": "on",
        "query": "select * from sales where store='1' and IsHoliday='TRUE'",
        "connectionName": "snowflake-glue-conn",
        "sfDatabase": "databaseName",
        "sfSchema": "schemaName",
        "sfWarehouse": "warehouseName",
    }
)
```

Writing to Snowflake tables

Prerequisites: A Snowflake database you would like to write to. You will need a current or desired table name, *`tableName`*. You will need your Snowflake url *`snowflakeUrl`*, username *`snowflakeUser`* and password *`snowflakePassword`*. If your Snowflake user does not have a default namespace set, you will need the Snowflake database name, *`databaseName`* and the schema name *`schemaName`*. Additionally, if your Snowflake user does not have a default warehouse set, you will need a warehouse name *`warehouseName`*.

For example:

Additional Prerequisites: Complete the steps *To manage your connection credentials with AWS Glue* to configure *`snowflakeUrl`*, *`snowflakeUsername`* and *`snowflakePassword`*. To review these steps, see the section called “Configuring Snowflake”, the previous section. To select which **Additional network connection** to connect with, we will use the `connectionName` parameter.

```
glueContext.write_dynamic_frame.from_options(
    connection_type="snowflake",
    connection_options={
        "connectionName": "connectionName",
        "dbtable": "tableName",
        "sfDatabase": "databaseName",
        "sfSchema": "schemaName",
        "sfWarehouse": "warehouseName",
    },
)
```

Snowflake connection option reference

The Snowflake connection type takes the following connection options:

You can retrieve some of the parameters in this section from a Data Catalog connection (`sfUrl`, `sfUser`, `sfPassword`), in which case you are not required to provide them. You can do this by providing the parameter `connectionName`.

You can retrieve connection parameters from AWS Secrets Manager secrets using the `secretId` parameter. When using Secrets Manager, the following Spark properties can be automatically retrieved if present in the secret:

- `sfUser` (using key `USERNAME` or `sfUser`)
- `sfPassword` (using key `PASSWORD` or `sfPassword`)
- `sfWarehouse` (using key `sfWarehouse`)
- `sfDatabase` (using key `sfDatabase`)
- `sfSchema` (using key `sfSchema`)
- `sfRole` (using key `sfRole`)
- `pem_private_key` (using key `pem_private_key`)

Property Precedence Order: When the same property is specified in multiple locations, AWS Glue uses the following precedence order (highest to lowest):

1. Explicitly provided connection options in your job code
2. Data Catalog connection properties
3. AWS Secrets Manager secret values (when `secretId` is specified)
4. Snowflake user defaults

The following parameters are used generally when connecting to Snowflake.

- `sfDatabase` — Required if a user default is not set in Snowflake. Used for Read/Write. The database to use for the session after connecting.
- `sfSchema` — Required if a user default is not set in Snowflake. Used for Read/Write. The schema to use for the session after connecting.
- `sfWarehouse` — Required if a user default is not set in Snowflake. Used for Read/Write. The default virtual warehouse to use for the session after connecting.
- `sfRole` — Required if a user default is not set in Snowflake. Used for Read/Write. The default security role to use for the session after connecting.
- `sfUrl` — (Required) Used for Read/Write. Specifies the hostname for your account in the following format: `account_identifier.snowflakecomputing.com`. For more information about account identifiers, see Account Identifiers in the Snowflake documentation.
- `sfUser` — (Required) Used for Read/Write. Login name for the Snowflake user.
- `sfPassword` — (Required unless `pem_private_key` provided) Used for Read/Write. Password for the Snowflake user.
- `dbtable` — Required when working with full tables. Used for Read/Write. The name of the table to be read or the table to which data is written. When reading, all columns and records are retrieved.
- `pem_private_key` — Used for Read/Write. An unencrypted b64-encoded private key string. The private key for the Snowflake user. It is common to copy this out of a PEM file. For more information, see Key-pair authentication and key-pair rotation in the Snowflake documentation.
- `query` — Required when reading with a query. Used for Read. The exact query (`SELECT` statement) to run

The following options are used to configure specific behaviors during the process of connecting to Snowflake.

- `preactions` — Used for Read/Write. Valid Values: Semicolon separated list of SQL statements as String. SQL statements run before data is transferred between AWS Glue and Snowflake. If a statement contains `%s`, the `%s` is replaced with the table name referenced for the operation.
- `postactions` — Used for Read/Write. SQL statements run after data is transferred between AWS Glue and Snowflake. If a statement contains `%s`, the `%s` is replaced with the table name referenced for the operation.
- `autopushdown` — Default: `"on"`. Valid Values: `"on"`, `"off"`. This parameter controls whether automatic query pushdown is enabled. If pushdown is enabled, then when a query is run on

Spark, if part of the query can be "pushed down" to the Snowflake server, it is pushed down. This improves performance of some queries. For information about whether your query can be pushed down, consult Pushdown in the Snowflake documentation.

Additionally, some of the options available on the Snowflake Spark connector may be supported in AWS Glue. For more information about options available on the Snowflake Spark connector, see Setting Configuration Options for the Connector in the Snowflake documentation.

Snowflake authentication methods

AWS Glue supports the following authentication methods for connecting to Snowflake:

- **Username and Password Authentication:** Provide `sfUser` and `sfPassword` parameters.
- **Key-Pair Authentication:** Provide `sfUser` and `pem_private_key` parameters. When using key-pair authentication, the `sfPassword` parameter is not required.

Both authentication methods are fully supported and can be configured using any combination of connection options, Data Catalog connections, or AWS Secrets Manager secrets.

Snowflake connector limitations

Connecting to Snowflake with AWS Glue for Spark is subject to the following limitations.

- This connector does not support job bookmarks. For more information about job bookmarks, see the section called “Tracking processed data using job bookmarks”.
- This connector does not support Snowflake reads and writes through tables in the AWS Glue Data Catalog using the `create_dynamic_frame.from_catalog` and `write_dynamic_frame.from_catalog` methods.
- This connector supports username/password authentication and key-pair authentication. Other authentication methods (such as OAuth or SAML) are not currently supported.
- This connector is not supported within streaming jobs.
- This connector supports `SELECT` statement based queries when retrieving information (such as with the `query` parameter). Other kind of queries (such as `SHOW`, `DESC`, or DML statements) are not supported.
- Snowflake limits the size of query text (i.e. SQL statements) submitted through Snowflake clients to 1 MB per statement. For more details, see Limits on Query Text Size.

Teradata Vantage connections

You can use AWS Glue for Spark to read from and write to existing tables in Teradata Vantage in AWS Glue 4.0 and later versions. You can define what to read from Teradata with a SQL query. You can connect to Teradata using username and password credentials stored in AWS Secrets Manager through a AWS Glue connection.

For more information about Teradata, consult the Teradata documentation

Configuring Teradata connections

To connect to Teradata from AWS Glue, you will need to create and store your Teradata credentials in an AWS Secrets Manager secret, then associate that secret with a AWS Glue Teradata connection. If your Teradata instance is in an Amazon VPC, you will also need to provide networking options to your AWS Glue Teradata connection.

To connect to Teradata from AWS Glue, you may need some prerequisites:

- If you are accessing your Teradata environment through Amazon VPC, configure Amazon VPC to allow your AWS Glue job to communicate with the Teradata environment. We discourage accessing the Teradata environment over the public internet.

  In Amazon VPC, identify or create a **VPC**, **Subnet** and **Security group** that AWS Glue will use while executing the job. Additionally, you need to ensure Amazon VPC is configured to permit network traffic between your Teradata instance and this location. Your job will need to establish a TCP connection with your Teradata client port. For more information about Teradata ports, see the Teradata documentation.

  Based on your network layout, secure VPC connectivity may require changes in Amazon VPC and other networking services. For more information about AWS connectivity, consult AWS Connectivity Options in the Teradata documentation.

To configure a AWS Glue Teradata connection:

1. In your Teradata configuration, identify or create a user and password AWS Glue will connect with, *`teradataUser`* and *`teradataPassword`*. For more information, consult Vantage Security Overview in the Teradata documentation.
2. In AWS Secrets Manager, create a secret using your Teradata credentials. To create a secret in Secrets Manager, follow the tutorial available in Create an AWS Secrets Manager secret in

the AWS Secrets Manager documentation. After creating the secret, keep the Secret name, *secretName* for the next step.

- When selecting **Key/value pairs**, create a pair for the key `user` with the value *teradataUsername*.
- When selecting **Key/value pairs**, create a pair for the key `password` with the value *teradataPassword*.

3. In the AWS Glue console, create a connection by following the steps in [the section called "Adding an AWS Glue connection"](). After creating the connection, keep the connection name, *connectionName*, for the next step.

   - When selecting a **Connection type**, select Teradata.
   - When providing **JDBC URL**, provide the URL for your instance. You can also hardcode certain comma separated connection parameters in your JDBC URL. The URL must conform to the following format: `jdbc:teradata://teradataHostname/ParameterName=ParameterValue,ParameterName`

     Supported URL parameters include:

     - `DATABASE`– name of database on host to access by default.
     - `DBS_PORT`– the database port, used when running on a nonstandard port.
   - When selecting a **Credential type**, select **AWS Secrets Manager**, then set **AWS Secret** to *secretName*.

4. In the following situations, you may require additional configuration:

   - For Teradata instances hosted on AWS in an Amazon VPC
     - You will need to provide Amazon VPC connection information to the AWS Glue connection that defines your Teradata security credentials. When creating or updating your connection, set **VPC**, **Subnet** and **Security groups** in **Network options**.

After creating a AWS Glue Teradata connection, you will need to perform the following steps before calling your connection method.

- Grant the IAM role associated with your AWS Glue job permission to read *secretName*.
- In your AWS Glue job configuration, provide *connectionName* as an **Additional network connection**.

Reading from Teradata

Prerequisites:

- A Teradata table you would like to read from. You will need the table name, *`tableName`*.
- A AWS Glue Teradata connection configured to provide auth information. Complete the steps *To configure a connection to Teradata* to configure your auth information. You will need the name of the AWS Glue connection, *`connectionName`*.

For example:

```
teradata_read_table = glueContext.create_dynamic_frame.from_options(
    connection_type="teradata",
    connection_options={
        "connectionName": "connectionName",
        "dbtable": "tableName"
    }
)
```

You can also provide a SELECT SQL query, to filter the results returned to your DynamicFrame. You will need to configure `query`.

For example:

```
teradata_read_query = glueContext.create_dynamic_frame.from_options(
    connection_type="teradata",
    connection_options={
        "connectionName": "connectionName",
        "query": "query"
    }
)
```

Writing to Teradata tables

Prerequisites: A Teradata table you would like to write to, *`tableName`*. **You must create the table before calling the connection method.**

For example:

```
teradata_write = glueContext.write_dynamic_frame.from_options(
    connection_type="teradata",
    connection_options={
        "connectionName": "connectionName",
        "dbtable": "tableName"
    }
)
```

Teradata connection option reference

- `connectionName` — Required. Used for Read/Write. The name of a AWS Glue Teradata connection configured to provide auth and networking information to your connection method.
- `dbtable` — Required for writing, required for reading unless `query` is provided. Used for Read/Write. The name of a table your connection method will interact with.
- `query` — Used for Read. A SELECT SQL query defining what should be retrieved when reading from Teradata.

Teradata Vantage NOS connections

Teradata NOS (Native Object Store) connection is a new connection for Teradata Vantage which leverages Teradata WRITE_NOS query to read from existing tables and READ_NOS query to write to tables. These queries uses Amazon S3 as a staging directory, and therefore the Teradata NOS connector is faster than the existing Teradata connector (JDBC-based) especially in handling large amount of data.

You can use the Teradata NOS connection in AWS Glue for Spark to read from and write to existing tables in Teradata Vantage in AWS Glue 5.0 and later versions. You can define what to read from Teradata with a SQL query. You can connect to Teradata using username and password credentials stored in AWS Secrets Manager through a AWS Glue connection.

For more information about Teradata, consult the Teradata documentation.

Topics

- Creating a Teradata NOS connection
- Reading from Teradata tables
- Writing to Teradata tables
- Teradata connection option reference
- Provide Options in AWS Glue Visual ETL UI

Creating a Teradata NOS connection

To connect to Teradata NOS from AWS Glue, you will need to create and store your Teradata credentials in an AWS Secrets Manager secret, then associate that secret with a AWS Glue Teradata NOS connection. If your Teradata instance is in an Amazon VPC, you will also need to provide networking options to your AWS Glue Teradata NOS connection.

Prerequisites:

- If you are accessing your Teradata environment through Amazon VPC, configure Amazon VPC to allow your AWS Glue job to communicate with the Teradata environment. We discourage accessing the Teradata environment over the public internet.
- In Amazon VPC, identify or create a VPC, Subnet and Security group that AWS Glue will use while executing the job. Additionally, you need to ensure Amazon VPC is configured to permit network traffic between your Teradata instance and this location. Your job will need to establish a TCP connection with your Teradata client port. For more information about Teradata ports, see the Security Groups for Teradata Vantage.
- Based on your network layout, secure VPC connectivity may require changes in Amazon VPC and other networking services. For more information about AWS connectivity, see AWS Connectivity Options in the Teradata documentation.

To configure a AWS Glue Teradata NOS connection:

1. In your Teradata configuration, identify or create a *teradataUsername* and *teradataPassword* AWS Glue will connect with. For more information, see Vantage Security Overview in the Teradata documentation.
2. In AWS Secrets Manager, create a secret using your Teradata credentials. To create a secret in AWS Secrets Manager, follow the tutorial available in Create an AWS Secrets Manager secret in the AWS Secrets Manager documentation. After creating the secret, keep the Secret name, *secretName* for the next step.

    - When selecting Key/value pairs, create a pair for the key USERNAME with the value *teradataUsername*.
    - When selecting Key/value pairs, create a pair for the key PASSWORD with the value *teradataPassword*.

3. In the AWS Glue console, create a connection by following the steps in [Adding an AWS Glue connection](). After creating the connection, keep the connection name, *`connectionName`*, for the next step.

   - When selecting a **Connection type**, select Teradata Vantage NOS.
   - When providing JDBC URL, provide the URL for your instance. You can also hardcode certain comma separated connection parameters in your JDBC URL. The URL must conform to the following format: `jdbc:teradata://teradataHostname/ParameterName=ParameterValue,ParameterName=ParameterValue` .
   - Supported URL parameters include:
     - `DATABASE`– name of database on host to access by default.
     - `DBS_PORT`– the database port, used when running on a nonstandard port.
   - When selecting a **Credential type**, select AWS Secrets Manager, then set **AWS Secret** to *`secretName`*.

4. In the following situations, you may require additional configuration:

   - For Teradata instances hosted on AWS in an Amazon VPC, you will need to provide Amazon VPC connection information to the AWS Glue connection that defines your Teradata security credentials. When creating or updating your connection, set **VPC**, **Subnet**, and **Security groups** in **Network options**.

After creating a AWS Glue Teradata Vantage NOS connection, you will need to perform the following steps before calling your connection method.

1. Grant the IAM role associated with your AWS Glue job permission to read *`secretName`*.
2. In your AWS Glue job configuration, provide *`connectionName`* as an **Additional network connection** Under **Connections**.

Reading from Teradata tables

Prerequisites:

- A Teradata table you would like to read from. You will need the table name, *`tableName`*.
- The Teradata environment has write access to the Amazon S3 path specified by `staging_fs_url` option, *`stagingFsUrl`*.

- The IAM role associated with AWS Glue job has write access to the Amazon S3 location specified by `staging_fs_url` option.
- An AWS Glue Teradata NOS connection configured to provide auth information. Complete the steps To configure a AWS Glue Teradata NOS connection: to configure your auth information. You will need the name of the AWS Glue connection, *connectionName*.

Example:

```
teradata_read_table = glueContext.create_dynamic_frame.from_options(
    connection_type= "teradatanos",
    connection_options={
        "connectionName": "connectionName",
        "dbtable": "tableName",
        "staging_fs_url": "stagingFsUrl"
    }
)
```

You can also provide a SELECT SQL query, to filter the results returned to your DynamicFrame. You will need to configure query. If you configure both dbTable and query, the connector fails to read the data. For example:

```
teradata_read_query = glueContext.create_dynamic_frame.from_options(
    connection_type="teradatanos",
    connection_options={
        "connectionName": "connectionName",
        "query": "query",
        "staging_fs_url": "stagingFsUrl"
    }
)
```

Additionally, you can use Spark DataFrame API to read from Teradata tables. For example:

```
options = {
    "url": "JDBC_URL",
    "dbtable": "tableName",
    "user": "teradataUsername", # or use "username" as key here
    "password": "teradataPassword",
    "staging_fs_url": "stagingFsUrl"
}
teradata_read_table = spark.read.format("teradatanos").option(**options).load()
```

Writing to Teradata tables

Prerequisites

- A Teradata table you would like to write to: *`tableName`*.
- The Teradata environment has read access to the Amazon S3 location specified by `staging_fs_url` option, *`stagingFsUrl`* .
- The IAM role associated with AWS Glue job has write access to the Amazon S3 location specified by `staging_fs_url` option.
- An AWS Glue Teradata connection configured to provide auth information. Complete the steps in To configure a AWS Glue Teradata NOS connection: to configure your auth information. You will need the name of the AWS Glue connection, *`connectionName`*.

  For example:

  ```
  teradata_write = glueContext.write_dynamic_frame.from_options(
      frame=dynamicFrame,
      connection_type= "teradatanos",
      connection_options={
          "connectionName": "connectionName",
          "dbtable": "tableName",
          "staging_fs_url": "stagingFsUrl"
      }
  )
  ```

Teradata connection option reference

Connection and Operation Options:

- `connectionName` — Required. Used for Read/Write. The name of an AWS Glue Teradata connection configured to provide auth and networking information to your connection method.
- `staging_fs_url` — Required. Used for Read/Write. A writable location in Amazon S3, to be used for unloaded data when reading from Teradata, and for Parquet data to be loaded into Redshift when writing to Teradata. The S3 bucket must be in the same region as the region of your AWS Glue jobs.
- `dbtable` — Required for writing, required for reading unless `query` is provided. Used for Read/Write. The name of a table your connection method will interact with.

- `query` — Used for Read. A SELECT SQL query defining what should be retrieved when reading from Teradata. You can not pass if `dbtable` option is provided.
- `clean_staging_s3_dir` — Optional. Used for Read/Write. If true, clean up the staging Amazon S3 objects after a read or a write. The default value is true.
- `pre_actions` — Optional. Used for Write. Semicolon-separated list of SQL commands that are executed before data is transferred between Spark and Teradata Vantage.
- `post_actions` — Optional. Used for Write. Semicolon-separated list of SQL commands that are executed after data is transferred between Spark and Teradata Vantage.
- `truncate` — Optional. Used for Write. If true, the connector truncates the table when writing in overwrite mode. If false, the connector drops the table when writing in overwrite mode. The default value is false.
- `create_table_script` — Optional. Used for Write. An SQL statement to create table when writing to Teradata Vantage. Useful when you want to create a table with custom metadata (e.g. CREATE MULTISET or SET table or change primary index). Note that the table name used in create table script should match with the table name specified in `dbtable` option.
- `partition_size_in_mb` — Optional. Used for Read. Maximum size of a Spark partition in megabytes while reading staging Amazon S3 objects. The default value is 128.

You can provide advanced options when creating a Teradata node. These options are the same as those available when programming AWS Glue for Spark scripts.

See the section called “Teradata Vantage connections”.

Authorization Options:

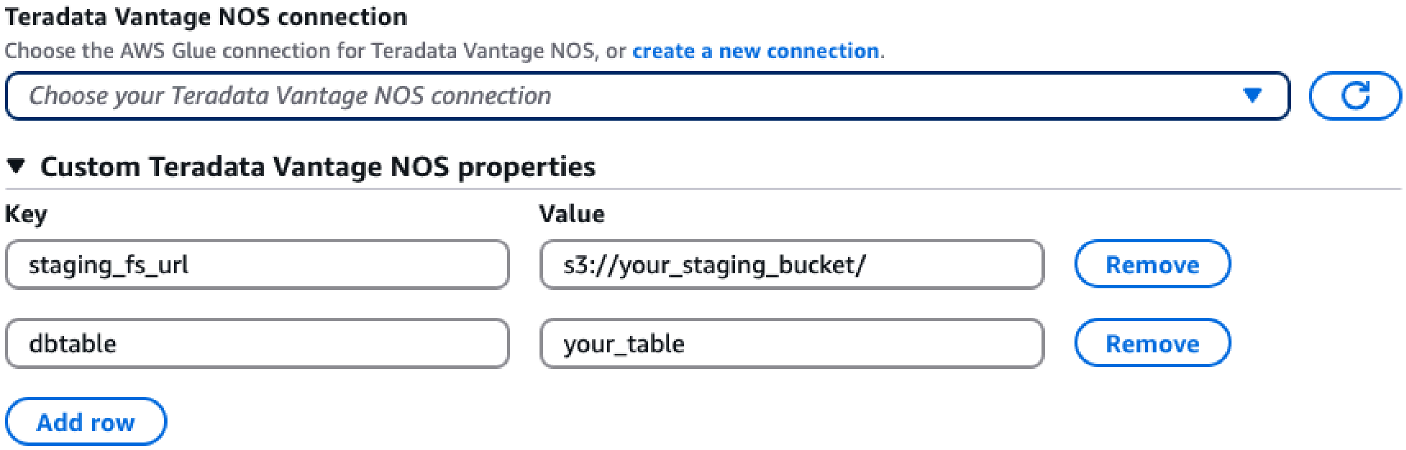
Below are options used to provide AWS account credentials that the connector uses to access the staging Amazon S3 bucket. You can choose to (1) not provide any authorization options at all, and use temporary credentials generated from your AWS Glue execution role; or (2) provide an authorization object, `auth_object` you created; or (3) provide `aws_access_key_id` and `aws_secret_access_key` if using long term credentials, or provide `aws_access_key`, `aws_secret_access_key`, and `aws_session_token` if using temporary credentials.

- `auth_object` – Optional. Used for accessing the staging Amazon S3 bucket. An authorization object string created in Teradata instance. If provided, the connector will use this authorization object to access the staging Amazon S3 bucket. If not provided, and `aws_access_key_id` and `aws_secret_access_key` are also not provided, a temporary credential will be retrieved from AWS Glue execution role and used by the connector. The AWS account associated with this authorization object must be in the same region as your AWS Glue jobs and your staging Amazon S3 bucket or configured with cross account trust.

- `aws_access_key_id` – Optional. Used for accessing the staging Amazon S3 bucket. Part of an AWS account security credential. If `auth_object` is not provided, and `aws_access_key_id` is provided with `aws_secret_access_key`, the connector will use them to access staging Amazon S3 bucket. The AWS account associated with this access key must be in the same region as your AWS Glue jobs and your staging Amazon S3 bucket or configured with cross account trust.
- `aws_secret_access_key` – Optional. Used for accessing the staging Amazon S3 bucket. Part of an AWS account security credential. If `auth_object` is not provided, and `aws_secret_access_key` is provided with `aws_access_key_id` , the connector will use them to access staging Amazon S3 bucket. The AWS account associated with this secret key must be in the same region as your AWS Glue jobs and your staging Amazon S3 bucket or configured with cross account trust.
- `aws_session_token` – Optional. Used for accessing the staging Amazon S3 bucket. Part of a temporary AWS account security credential. Should be provided with `aws_access_key_id` and `aws_secret_access_key`.

Provide Options in AWS Glue Visual ETL UI

You can provide all above options in your visual ETL job UI. For connectionName option, you should choose it from the Teradata Vantage NOS connection drop down list. For all other options, you should provide them through the Custom Teradata Vantage NOS properties as key value pairs.



Teradata Vantage NOS connection

Choose the AWS Glue connection for Teradata Vantage NOS, or create a new connection.

Choose your Teradata Vantage NOS connection

Custom Teradata Vantage NOS properties

Key	Value
staging_fs_url	s3://your_staging_bucket/
dbtable	your_table

Vertica connections

You can use AWS Glue for Spark to read from and write to tables in Vertica in AWS Glue 4.0 and later versions. You can define what to read from Vertica with a SQL query. You connect to Vertica using username and password credentials stored in AWS Secrets Manager through a AWS Glue connection.

For more information about Vertica, consult the Vertica documentation.

Configuring Vertica connections

To connect to Vertica from AWS Glue, you will need to create and store your Vertica credentials in a AWS Secrets Manager secret, then associate that secret with a Vertica AWS Glue connection. If your Vertica instance is in an Amazon VPC, you will also need to provide networking options to your AWS Glue Vertica connection. You will need an Amazon S3 bucket or folder to use for temporary storage when reading from and writing to the database.

To connect to Vertica from AWS Glue, you will need some prerequisites:

- An Amazon S3 bucket or folder to use for temporary storage when reading from and writing to the database, referred to by *tempS3Path*.

  > **Note**
  >
  > When using Vertica in AWS Glue job data previews, temporary files may not be automatically removed from *tempS3Path*. To ensure the removal of temporary files, directly end the data preview session by choosing **End session** in the **Data preview** pane. If you cannot guarantee the data preview session is ended directly, consider setting Amazon S3 Lifecycle configuration to remove old data. We recommend removing data older than 49 hours, based on maximum job runtime plus a margin. For more information about configuring Amazon S3 Lifecycle, see Managing your storage lifecycle in the Amazon S3 documentation.

- An IAM policy with appropriate permissions to your Amazon S3 path you can associate with your AWS Glue job role.
- If your Vertica instance is in an Amazon VPC, configure Amazon VPC to allow your AWS Glue job to communicate with the Vertica instance without traffic traversing the public internet.

  In Amazon VPC, identify or create a **VPC**, **Subnet** and **Security group** that AWS Glue will use while executing the job. Additionally, you need to ensure Amazon VPC is configured to permit network traffic between your Vertica instance and this location. Your job will need to establish a TCP connection with your Vertica client port, (default 5433). Based on your network layout, this may require changes to security group rules, Network ACLs, NAT Gateways and Peering connections.

You can then proceed to configure AWS Glue for use with Vertica.

To configure a connection to Vertica:

1. In AWS Secrets Manager, create a secret using your Vertica credentials, *`verticaUsername`* and *`verticaPassword`*. To create a secret in Secrets Manager, follow the tutorial available in Create an AWS Secrets Manager secret in the AWS Secrets Manager documentation. After creating the secret, keep the Secret name, *`secretName`* for the next step.
    - When selecting **Key/value pairs**, create a pair for the key `user` with the value *`verticaUsername`*.
    - When selecting **Key/value pairs**, create a pair for the key `password` with the value *`verticaPassword`*.
2. In the AWS Glue console, create a connection by following the steps in the section called “Adding an AWS Glue connection”. After creating the connection, keep the connection name, *`connectionName`*, for the next step.
    - When selecting a **Connection type**, select Vertica.
    - When selecting **Vertica Host**, provide the hostname of your Vertica installation.
    - When selecting **Vertica Port**, the port your Vertica installation is available through.
    - When selecting an **AWS Secret**, provide *`secretName`*.
3. In the following situations, you may require additional configuration:
    - For Vertica instances hosted on AWS in an Amazon VPC
        - Provide Amazon VPC connection information to the AWS Glue connection that defines your Vertica security credentials. When creating or updating your connection, set **VPC**, **Subnet** and **Security groups** in **Network options**.

After creating a AWS Glue Vertica connection, you will need to perform the following steps before calling your connection method.

- Grant the IAM role associated with your AWS Glue job permissions to *`tempS3Path`*.
- Grant the IAM role associated with your AWS Glue job permission to read *`secretName`*.
- In your AWS Glue job configuration, provide *`connectionName`* as an **Additional network connection**.

Reading from Vertica

Prerequisites:

- A Vertica table you would like to read from. You will need the Vertica database name, *`dbName`* and the table name, *`tableName`*.
- A AWS Glue Vertica connection configured to provide auth information. Complete the steps in the previous procedure, *To configure a connection to Vertica* to configure your auth information. You will need the name of the AWS Glue connection, *`connectionName`*.
- A Amazon S3 bucket or folder to use for temporary storage, mentioned previously. You will need the name, *`tempS3Path`*. You will need to connect to this location using the `s3a` protocol.

For example:

```
dynamicFrame = glueContext.create_dynamic_frame.from_options(
    connection_type="vertica",
    connection_options={
        "connectionName": "connectionName",
        "staging_fs_url": "s3a://tempS3Path",
        "db": "dbName",
        "table": "tableName",
    }
)
```

You can also provide a SELECT SQL query, to filter the results returned to your DynamicFrame or to access a dataset from multiple tables.

For example:

```
dynamicFrame = glueContext.create_dynamic_frame.from_options(
    connection_type="vertica",
    connection_options={
        "connectionName": "connectionName",
        "staging_fs_url": "s3a://tempS3Path",
        "db": "dbName",
        "query": "select * FROM tableName",
    },
)
```

Writing to Vertica tables

This example writes information from an existing DynamicFrame, *`dynamicFrame`* to Vertica. If the table already has information, AWS Glue will append data from your DynamicFrame.

Prerequisites:

- A current or desired table name, *`tableName`*, you would like to write to. You will also need the corresponding Vertica database name, *`dbName`*.
- A AWS Glue Vertica connection configured to provide auth information. Complete the steps in the previous procedure, *To configure a connection to Vertica* to configure your auth information. You will need the name of the AWS Glue connection, *`connectionName`*.
- A Amazon S3 bucket or folder to use for temporary storage, mentioned previously. You will need the name, *`tempS3Path`*. You will need to connect to this location using the `s3a` protocol.

For example:

```
glueContext.write_dynamic_frame.from_options(
    frame=dynamicFrame,
    connection_type="vertica",
    connection_options={
        "connectionName": "connectionName",
        "staging_fs_url": "s3a://tempS3Path",
        "db": "dbName",
        "table": "tableName",
    }
)
```

Vertica connection option reference

- `connectionName` — Required. Used for Read/Write. The name of a AWS Glue Vertica connection configured to provide auth and networking information to your connection method.
- `db` — Required. Used for Read/Write. The name of a database in Vertica your connection method will interact with.

- `dbSchema` — Required if needed to identify your table. Used for Read/Write. Default: `public`. The name of a schema your connection method will interact with.
- `table` — Required for writing, required for reading unless `query` is provided. Used for Read/Write. The name of a table your connection method will interact with.
- `query` — Used for Read. A SELECT SQL query defining what should be retrieved when reading from Teradata.
- `staging_fs_url` — Required. Used for Read/Write. Valid Values: `s3a` URLs. The URL of a Amazon S3 bucket or folder to use for temporary storage.

DataFrame options for ETL in AWS Glue 5.0 for Spark

A DataFrame is a Dataset organized into named columns similar to a table and supports functional-style (map/reduce/filter/etc.) operations and SQL operations (select, project, aggregate).

To create a DataFrame for a data source supported by Glue, the following are required:

- data source connector `ClassName`
- data source connection `Options`

Similarly, to write a DataFrame to a data sink supported by Glue, the same are required:

- data sink connector `ClassName`
- data sink connection `Options`

Note that AWS Glue features such as job bookmarks and DynamicFrame options such as `connectionName` are not supported in DataFrame. For more details about DataFrame and the supported operations, see the Spark documentation for DataFrame.

Specifying the connector ClassName

To specify the `ClassName` of a data source/sink, use the `.format` option to provide the corresponding connector `ClassName` that defines the data source/sink.

JDBC connectors

For JDBC connectors, specify `jdbc` as the value of the `.format` option and provide the JDBC driver `ClassName` in the `driver` option.

```
df = spark.read.format("jdbc").option("driver", "<DATA SOURCE JDBC DRIVER
 CLASSNAME>")...

df.write.format("jdbc").option("driver", "<DATA SINK JDBC DRIVER CLASSNAME>")...
```

The following table lists the JDBC driver `ClassName` of the supported data source in AWS Glue for DataFrames.

Data source	Driver ClassName
PostgreSQL	org.postgresql.Driver
Oracle	oracle.jdbc.driver.OracleDriver
SQLServer	com.microsoft.sqlserver.jdbc.SQLServerDriver
MySQL	com.mysql.jdbc.Driver
SAPHana	com.sap.db.jdbc.Driver
Teradata	com.teradata.jdbc.TeraDriver

Spark connectors

For spark connectors, specify the `ClassName` of the connector as the value of the `.format` option.

```
df = spark.read.format("<DATA SOURCE CONNECTOR CLASSNAME>")...

df.write.format("<DATA SINK CONNECTOR CLASSNAME>")...
```

The following table lists the Spark connector `ClassName` of the supported data source in AWS Glue for DataFrames.

Data source	ClassName
MongoDB/DocumentDB	glue.spark.mongodb
Redshift	io.github.spark_redshift_community.spark.redshift

AzureCosmos	cosmos.oltp
AzureSQL	com.microsoft.sqlserver.jdbc.spark
BigQuery	com.google.cloud.spark.bigquery
OpenSearch	org.opensearch.spark.sql
Snowflake	net.snowflake.spark.snowflake
Vertica	com.vertica.spark.datasource.VerticaSource

Specifying the connection Options

To specify the `Options` of the connection to a data source/sink, use the `.option(<KEY>, <VALUE>)` to provide individual options or `.options(<MAP>)` to provide multiple options as a key-value map.

Each data source/sink supports its own set of connection `Options`. For details on the available `Options`, refer to the public documentation of the specific data source/sink Spark connector listed in the following table.

- JDBC
- MongoDB/DocumentDB
- Redshift
- AzureCosmos
- AzureSQL
- BigQuery
- OpenSearch
- Snowflake
- Vertica

Examples

The following examples read from PostgreSQL and write into SnowFlake:

Python

Example:

```
from awsglue.context import GlueContext
from pyspark.sql import SparkSession

spark = SparkSession.builder.getOrCreate()

dataSourceClassName = "jdbc"
dataSourceOptions = {
  "driver": "org.postgresql.Driver",
  "url": "<url>",
  "user": "<user>",
  "password": "<password>",
  "dbtable": "<dbtable>",
}

dataframe = spark.read.format(className).options(**options).load()

dataSinkClassName = "net.snowflake.spark.snowflake"
dataSinkOptions = {
  "sfUrl": "<url>",
  "sfUsername": "<username>",
  "sfPassword": "<password>",
  "sfDatabase" -> "<database>",
  "sfSchema" -> "<schema>",
  "sfWarehouse" -> "<warehouse>"
}

dataframe.write.format(dataSinkClassName).options(**dataSinkOptions).save()
```

Scala

Example:

```
import org.apache.spark.sql.SparkSession

val spark = SparkSession.builder().getOrCreate()

val dataSourceClassName = "jdbc"
val dataSourceOptions = Map(
  "driver" -> "org.postgresql.Driver",
  "url" -> "<url>",
  "user" -> "<user>",
  "password" -> "<password>",
  "dbtable" -> "<dbtable>"
)

val dataframe =
 spark.read.format(dataSourceClassName).options(dataSourceOptions).load()

val dataSinkClassName = "net.snowflake.spark.snowflake"
val dataSinkOptions = Map(
  "sfUrl" -> "<url>",
  "sfUsername" -> "<username>",
  "sfPassword" -> "<password>",
  "sfDatabase" -> "<database>",
  "sfSchema" -> "<schema>",
  "sfWarehouse" -> "<warehouse>"
)

dataframe.write.format(dataSinkClassName).options(dataSinkOptions).save()
```

Custom and AWS Marketplace connectionType values

These include the following:

- `"connectionType": "marketplace.athena"`: Designates a connection to an Amazon Athena data store. The connection uses a connector from AWS Marketplace.
- `"connectionType": "marketplace.spark"`: Designates a connection to an Apache Spark data store. The connection uses a connector from AWS Marketplace.
- `"connectionType": "marketplace.jdbc"`: Designates a connection to a JDBC data store. The connection uses a connector from AWS Marketplace.
- `"connectionType": "custom.athena"`: Designates a connection to an Amazon Athena data store. The connection uses a custom connector that you upload to AWS Glue Studio.
- `"connectionType": "custom.spark"`: Designates a connection to an Apache Spark data store. The connection uses a custom connector that you upload to AWS Glue Studio.
- `"connectionType": "custom.jdbc"`: Designates a connection to a JDBC data store. The connection uses a custom connector that you upload to AWS Glue Studio.

Connection options for type custom.jdbc or marketplace.jdbc

- `className` – String, required, driver class name.

- `connectionName` – String, required, name of the connection that is associated with the connector.
- `url` – String, required, JDBC URL with placeholders (`${}`) which are used to build the connection to the data source. The placeholder `${secretKey}` is replaced with the secret of the same name in AWS Secrets Manager. Refer to the data store documentation for more information about constructing the URL.
- `secretId` or `user/password` – String, required, used to retrieve credentials for the URL.
- `dbTable` or `query` – String, required, the table or SQL query to get the data from. You can specify either `dbTable` or `query`, but not both.
- `partitionColumn` – String, optional, the name of an integer column that is used for partitioning. This option works only when it's included with `lowerBound`, `upperBound`, and `numPartitions`. This option works the same way as in the Spark SQL JDBC reader. For more information, see JDBC To Other Databases in the *Apache Spark SQL, DataFrames and Datasets Guide*.

  The `lowerBound` and `upperBound` values are used to decide the partition stride, not for filtering the rows in table. All rows in the table are partitioned and returned.

  > **Note**
  >
  > When using a query instead of a table name, you should validate that the query works with the specified partitioning condition. For example:
  >
  > - If your query format is `"SELECT col1 FROM table1"`, then test the query by appending a `WHERE` clause at the end of the query that uses the partition column.
  > - If your query format is "`SELECT col1 FROM table1 WHERE col2=val`", then test the query by extending the `WHERE` clause with `AND` and an expression that uses the partition column.

- `lowerBound` – Integer, optional, the minimum value of `partitionColumn` that is used to decide partition stride.
- `upperBound` – Integer, optional, the maximum value of `partitionColumn` that is used to decide partition stride.
- `numPartitions` – Integer, optional, the number of partitions. This value, along with `lowerBound` (inclusive) and `upperBound` (exclusive), form partition strides for generated `WHERE` clause expressions that are used to split the `partitionColumn`.

> **Important**
>
> Be careful with the number of partitions because too many partitions might cause problems on your external database systems.

- `filterPredicate` – String, optional, extra condition clause to filter data from source. For example:

  ```
  BillingCity='Mountain View'
  ```

  When using a *query* instead of a *table* name, you should validate that the query works with the specified `filterPredicate`. For example:

  - If your query format is `"SELECT col1 FROM table1"`, then test the query by appending a `WHERE` clause at the end of the query that uses the filter predicate.
  - If your query format is `"SELECT col1 FROM table1 WHERE col2=val"`, then test the query by extending the `WHERE` clause with `AND` and an expression that uses the filter predicate.
- `dataTypeMapping` – Dictionary, optional, custom data type mapping that builds a mapping from a **JDBC** data type to a **Glue** data type. For example, the option `"dataTypeMapping":{"FLOAT":"STRING"}` maps data fields of JDBC type `FLOAT` into the Java `String` type by calling the `ResultSet.getString()` method of the driver, and uses it to build AWS Glue records. The `ResultSet` object is implemented by each driver, so the behavior is specific to the driver you use. Refer to the documentation for your JDBC driver to understand how the driver performs the conversions.
- The AWS Glue data types supported currently are:
  - DATE
  - STRING
  - TIMESTAMP
  - INT
  - FLOAT
  - LONG
  - BIGDECIMAL
  - BYTE
  - SHORT

- DOUBLE

The JDBC data types supported are Java8 java.sql.types.

The default data type mappings (from JDBC to AWS Glue) are:

- DATE -> DATE
- VARCHAR -> STRING
- CHAR -> STRING
- LONGNVARCHAR -> STRING
- TIMESTAMP -> TIMESTAMP
- INTEGER -> INT
- FLOAT -> FLOAT
- REAL -> FLOAT
- BIT -> BOOLEAN
- BOOLEAN -> BOOLEAN
- BIGINT -> LONG
- DECIMAL -> BIGDECIMAL
- NUMERIC -> BIGDECIMAL
- TINYINT -> SHORT
- SMALLINT -> SHORT
- DOUBLE -> DOUBLE

If you use a custom data type mapping with the option `dataTypeMapping`, then you can override a default data type mapping. Only the JDBC data types listed in the `dataTypeMapping` option are affected; the default mapping is used for all other JDBC data types. You can add mappings for additional JDBC data types if needed. If a JDBC data type is not included in either the default mapping or a custom mapping, then the data type converts to the AWS Glue `STRING` data type by default.

The following Python code example shows how to read from JDBC databases with AWS Marketplace JDBC drivers. It demonstrates reading from a database and writing to an S3 location.

```
import sys
from awsglue.transforms import *
```

```
    from awsglue.utils import getResolvedOptions
    from pyspark.context import SparkContext
    from awsglue.context import GlueContext
    from awsglue.job import Job

    ## @params: [JOB_NAME]
    args = getResolvedOptions(sys.argv, ['JOB_NAME'])

    sc = SparkContext()
    glueContext = GlueContext(sc)
    spark = glueContext.spark_session
    job = Job(glueContext)
    job.init(args['JOB_NAME'], args)
    ## @type: DataSource
    ## @args: [connection_type = "marketplace.jdbc", connection_options =
     {"dataTypeMapping":{"INTEGER":"STRING"},"upperBound":"200","query":"select id,
       name, department from department where id < 200","numPartitions":"4",
       "partitionColumn":"id","lowerBound":"0","connectionName":"test-connection-
jdbc"},
        transformation_ctx = "DataSource0"]
    ## @return: DataSource0
    ## @inputs: []
    DataSource0 = glueContext.create_dynamic_frame.from_options(connection_type =
      "marketplace.jdbc", connection_options = {"dataTypeMapping":{"INTEGER":"STRING"},
      "upperBound":"200","query":"select id, name, department from department where
       id < 200","numPartitions":"4","partitionColumn":"id","lowerBound":"0",
       "connectionName":"test-connection-jdbc"}, transformation_ctx = "DataSource0")
    ## @type: ApplyMapping
    ## @args: [mappings = [("department", "string", "department", "string"), ("name",
 "string",
      "name", "string"), ("id", "int", "id", "int")], transformation_ctx =
 "Transform0"]
    ## @return: Transform0
    ## @inputs: [frame = DataSource0]
    Transform0 = ApplyMapping.apply(frame = DataSource0, mappings = [("department",
 "string",
      "department", "string"), ("name", "string", "name", "string"), ("id", "int",
 "id", "int")],
       transformation_ctx = "Transform0")
    ## @type: DataSink
    ## @args: [connection_type = "s3", format = "json", connection_options = {"path":
     "s3://<S3 path>/", "partitionKeys": []}, transformation_ctx = "DataSink0"]
    ## @return: DataSink0
    ## @inputs: [frame = Transform0]
```

```
DataSink0 = glueContext.write_dynamic_frame.from_options(frame = Transform0,
  connection_type = "s3", format = "json", connection_options = {"path":
  "s3://<S3 path>/", "partitionKeys": []}, transformation_ctx = "DataSink0")
job.commit()
```

Connection options for type custom.athena or marketplace.athena

- `className` – String, required, driver class name. When you're using the Athena-CloudWatch connector, this parameter value is the prefix of the class Name (for example, `"com.amazonaws.athena.connectors"`). The Athena-CloudWatch connector is composed of two classes: a metadata handler and a record handler. If you supply the common prefix here, then the API loads the correct classes based on that prefix.
- `tableName` – String, required, the name of the CloudWatch log stream to read. This code snippet uses the special view name `all_log_streams`, which means that the dynamic data frame returned will contain data from all log streams in the log group.
- `schemaName` – String, required, the name of the CloudWatch log group to read from. For example, `/aws-glue/jobs/output`.
- `connectionName` – String, required, name of the connection that is associated with the connector.

For additional options for this connector, see the Amazon Athena CloudWatch Connector README file on GitHub.

The following Python code example shows how to read from an Athena data store using an AWS Marketplace connector. It demonstrates reading from Athena and writing to an S3 location.

```
import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job

## @params: [JOB_NAME]
args = getResolvedOptions(sys.argv, ['JOB_NAME'])

sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
```

```
   job = Job(glueContext)
   job.init(args['JOB_NAME'], args)
   ## @type: DataSource
   ## @args: [connection_type = "marketplace.athena", connection_options =
    {"tableName":"all_log_streams","schemaName":"/aws-glue/jobs/output",
     "connectionName":"test-connection-athena"}, transformation_ctx = "DataSource0"]
   ## @return: DataSource0
   ## @inputs: []
   DataSource0 = glueContext.create_dynamic_frame.from_options(connection_type =
     "marketplace.athena", connection_options = {"tableName":"all_log_streams",,
     "schemaName":"/aws-glue/jobs/output","connectionName":
     "test-connection-athena"}, transformation_ctx = "DataSource0")
   ## @type: ApplyMapping
   ## @args: [mappings = [("department", "string", "department", "string"), ("name",
"string",
     "name", "string"), ("id", "int", "id", "int")], transformation_ctx =
"Transform0"]
   ## @return: Transform0
   ## @inputs: [frame = DataSource0]
   Transform0 = ApplyMapping.apply(frame = DataSource0, mappings = [("department",
"string",
     "department", "string"), ("name", "string", "name", "string"), ("id", "int",
"id", "int")],
      transformation_ctx = "Transform0")
   ## @type: DataSink
   ## @args: [connection_type = "s3", format = "json", connection_options = {"path":
    "s3://<S3 path>/", "partitionKeys": []}, transformation_ctx = "DataSink0"]
   ## @return: DataSink0
   ## @inputs: [frame = Transform0]
   DataSink0 = glueContext.write_dynamic_frame.from_options(frame = Transform0,
     connection_type = "s3", format = "json", connection_options = {"path":
     "s3://<S3 path>/", "partitionKeys": []}, transformation_ctx = "DataSink0")
   job.commit()
```

Connection options for type custom.spark or marketplace.spark

- `className` – String, required, connector class name.
- `secretId` – String, optional, used to retrieve credentials for the connector connection.
- `connectionName` – String, required, name of the connection that is associated with the connector.
- Other options depend on the data store. For example, OpenSearch configuration options start with the prefix `es`, as described in the Elasticsearch for Apache Hadoop documentation. Spark

connections to Snowflake use options such as `sfUser` and `sfPassword`, as described in Using the Spark Connector in the *Connecting to Snowflake* guide.

The following Python code example shows how to read from an OpenSearch data store using a `marketplace.spark` connection.

```
   import sys
   from awsglue.transforms import *
   from awsglue.utils import getResolvedOptions
   from pyspark.context import SparkContext
   from awsglue.context import GlueContext
   from awsglue.job import Job

   ## @params: [JOB_NAME]
   args = getResolvedOptions(sys.argv, ['JOB_NAME'])

   sc = SparkContext()
   glueContext = GlueContext(sc)
   spark = glueContext.spark_session
   job = Job(glueContext)
   job.init(args['JOB_NAME'], args)
   ## @type: DataSource
   ## @args: [connection_type = "marketplace.spark", connection_options = 
{"path":"test",
     "es.nodes.wan.only":"true","es.nodes":"https://<AWS endpoint>",
     "connectionName":"test-spark-es","es.port":"443"}, transformation_ctx = 
"DataSource0"]
   ## @return: DataSource0
   ## @inputs: []
   DataSource0 = glueContext.create_dynamic_frame.from_options(connection_type =
     "marketplace.spark", connection_options = {"path":"test","es.nodes.wan.only":
     "true","es.nodes":"https://<AWS endpoint>","connectionName":
     "test-spark-es","es.port":"443"}, transformation_ctx = "DataSource0")
   ## @type: DataSink
   ## @args: [connection_type = "s3", format = "json", connection_options = {"path":
        "s3://<S3 path>/", "partitionKeys": []}, transformation_ctx = "DataSink0"]
   ## @return: DataSink0
   ## @inputs: [frame = DataSource0]
   DataSink0 = glueContext.write_dynamic_frame.from_options(frame = DataSource0,
      connection_type = "s3", format = "json", connection_options = {"path":
      "s3://<S3 path>/", "partitionKeys": []}, transformation_ctx = "DataSink0")
   job.commit()
```

General options

The options in this section are provided as `connection_options`, but do not specifically apply to one connector.

The following parameters are used generally when configuring bookmarks. They may apply to Amazon S3 or JDBC workflows. For more information, see the section called "Using job bookmarks".

- `jobBookmarkKeys` — Array of column names.
- `jobBookmarkKeysSortOrder` — String defining how to compare values based on sort order. Valid values: `"asc"`, `"desc"`.
- `useS3ListImplementation` — Used to manage memory performance when listing Amazon S3 bucket contents. For more information, see Optimize memory management in AWS Glue.

Data format options for inputs and outputs in AWS Glue for Spark

These pages offer information about feature support and configuration parameters for data formats supported by AWS Glue for Spark. See the following for a description of the usage and applicablity of this information.

Feature support across data formats in AWS Glue

Each data format may support different AWS Glue features. The following common features may or may not be supported based on your format type. Refer to the documentation for your data format to understand how to leverage our features to meet your requirements.

Read	AWS Glue can recognize and interpret this data format without additional resources, such as connectors.
Write	AWS Glue can write data in this format without additional resources. You can include third-party libraries in your job and use standard Apache Spark functions to write data, as you would in other Spark environments. For more information about including

	libraries, see the section called "Python libraries".
Streamir read	AWS Glue can recognize and interpret this data format from an Apache Kafka, Amazon Managed Streaming for Apache Kafka or Amazon Kinesis message stream. We expect streams to present data in a consistent format, so they are read in as `DataFrames` .
Group small files	AWS Glue can group files together to batch work sent to each node when performing AWS Glue transforms. This can significantly improve performance for workloads involving large amounts of small files. For more information, see the section called "Grouping input files".
Job bookmai	AWS Glue can track the progress of transforms performing the same work on the same dataset across job runs with job bookmarks. This can improve performance for workloads involving datasets where work only needs to be done on new data since the last job run. For more information, see the section called "Tracking processed data using job bookmarks ".

Parameters used to interact with data formats in AWS Glue

Certain AWS Glue connection types support multiple `format` types, requiring you to specify information about your data format with a `format_options` object when using methods like `GlueContext.write_dynamic_frame.from_options`.

- `s3` – For more information, see Connection types and options for ETL in AWS Glue: S3 connection parameters. You can also view the documentation for the methods facilitating this connection type: the section called "create_dynamic_frame_from_options" and the section called "write_dynamic_frame_from_options" in Python and the corresponding Scala methods the section called "getSourceWithFormat" and the section called "getSinkWithFormat".

- `kinesis` – For more information, see Connection types and options for ETL in AWS Glue: Kinesis connection parameters. You can also view the documentation for the method facilitating this connection type: the section called "create_data_frame_from_options" and the corresponding Scala method the section called "createDataFrameFromOptions".
- `kafka` – For more information, see Connection types and options for ETL in AWS Glue: Kafka connection parameters. You can also view the documentation for the method facilitating this connection type: the section called "create_data_frame_from_options" and the corresponding Scala method the section called "createDataFrameFromOptions".

Some connection types do not require `format_options`. For example, in normal use, a JDBC connection to a relational database retrieves data in a consistent, tabular data format. Therefore, reading from a JDBC connection would not require `format_options`.

Some methods to read and write data in glue do not require `format_options`. For example, using `GlueContext.create_dynamic_frame.from_catalog` with AWS Glue crawlers. Crawlers determine the shape of your data. When using crawlers, a AWS Glue classifier will examine your data to make smart decisions about how to represent your data format. It will then store a representation of your data in the AWS Glue Data Catalog, which can be used within a AWS Glue ETL script to retrieve your data with the `GlueContext.create_dynamic_frame.from_catalog` method. Crawlers remove the need to manually specify information about your data format.

For jobs that access AWS Lake Formation governed tables, AWS Glue supports reading and writing all formats supported by Lake Formation governed tables. For the current list of supported formats for AWS Lake Formation governed tables, see Notes and Restrictions for Governed Tables in the *AWS Lake Formation Developer Guide*.

Note

For writing Apache Parquet, AWS Glue ETL only supports writing to a governed table by specifying an option for a custom Parquet writer type optimized for Dynamic Frames. When writing to a governed table with the `parquet` format, you should add the key `useGlueParquetWriter` with a value of `true` in the table parameters.

Topics

- Using the CSV format in AWS Glue
- Using the Parquet format in AWS Glue
- Using the XML format in AWS Glue
- Using the Avro format in AWS Glue
- Using the grokLog format in AWS Glue
- Using the Ion format in AWS Glue
- Using the JSON format in AWS Glue
- Using the ORC format in AWS Glue
- Using data lake frameworks with AWS Glue ETL jobs
- Shared configuration reference

Using the CSV format in AWS Glue

AWS Glue retrieves data from sources and writes data to targets stored and transported in various data formats. If your data is stored or transported in the CSV data format, this document introduces you available features for using your data in AWS Glue.

AWS Glue supports using the comma-separated value (CSV) format. This format is a minimal, row-based data format. CSVs often don't strictly conform to a standard, but you can refer to RFC 4180 and RFC 7111 for more information.

You can use AWS Glue to read CSVs from Amazon S3 and from streaming sources as well as write CSVs to Amazon S3. You can read and write `bzip` and `gzip` archives containing CSV files from S3. You configure compression behavior on the S3 connection parameters instead of in the configuration discussed on this page.

The following table shows which common AWS Glue features support the CSV format option.

Read	Write	Streaming read	Group small files	Job bookmarks
Supported	Supported	Supported	Supported	Supported

Example: Read CSV files or folders from S3

Prerequisites: You will need the S3 paths (`s3path`) to the CSV files or folders that you want to read.

Configuration: In your function options, specify `format="csv"`. In your `connection_options`, use the `paths` key to specify `s3path`. You can configure how the reader interacts with S3 in `connection_options`. For details, see Connection types and options for ETL in AWS Glue: S3 connection parameters. You can configure how the reader interprets CSV files in your `format_options`. For details, see CSV Configuration Reference.

The following AWS Glue ETL script shows the process of reading CSV files or folders from S3.

We provide a custom CSV reader with performance optimizations for common workflows through the `optimizePerformance` configuration key. To determine if this reader is right for your workload, see the section called “Using optimized CSV reader”.

Python

For this example, use the create_dynamic_frame.from_options method.

```
# Example: Read CSV from S3
# For show, we handle a CSV with a header row.  Set the withHeader option.
# Consider whether optimizePerformance is right for your workflow.

from pyspark.context import SparkContext
from awsglue.context import GlueContext

sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)
spark = glueContext.spark_session

dynamicFrame = glueContext.create_dynamic_frame.from_options(
    connection_type="s3",
    connection_options={"paths": ["s3://s3path"]},
    format="csv",
    format_options={
        "withHeader": True,
        # "optimizePerformance": True,
    },
)
```

You can also use DataFrames in a script (`pyspark.sql.DataFrame`).

```
dataFrame = spark.read\
    .format("csv")\
    .option("header", "true")\
    .load("s3://s3path")
```

Scala

For this example, use the getSourceWithFormat operation.

```
// Example: Read CSV from S3
// For show, we handle a CSV with a header row.  Set the withHeader option.
// Consider whether optimizePerformance is right for your workflow.

import com.amazonaws.services.glue.util.JsonOptions
import com.amazonaws.services.glue.{DynamicFrame, GlueContext}
import org.apache.spark.SparkContext

object GlueApp {
  def main(sysArgs: Array[String]): Unit = {
    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)

    val dynamicFrame = glueContext.getSourceWithFormat(
      formatOptions=JsonOptions("""{"withHeader": true}"""),
      connectionType="s3",
      format="csv",
      options=JsonOptions("""{"paths": ["s3://s3path"], "recurse": true}""")
    ).getDynamicFrame()
  }
}
```

You can also use DataFrames in a script (`org.apache.spark.sql.DataFrame`).

```
val dataFrame = spark.read
  .option("header","true")
  .format("csv")
  .load("s3://s3path“)
```

Example: Write CSV files and folders to S3

Prerequisites: You will need an initialized DataFrame (`dataFrame`) or a DynamicFrame (`dynamicFrame`). You will also need your expected S3 output path, `s3path`.

Configuration: In your function options, specify `format="csv"`. In your `connection_options`, use the `paths` key to specify `s3path`. You can configure how the writer interacts with S3 in `connection_options`. For details, see Connection types and options for ETL in AWS Glue: S3 connection parameters. You can configure how your operation writes the contents of your files in `format_options`. For details, see CSV Configuration Reference. The following AWS Glue ETL script shows the process of writing CSV files and folders to S3.

Python

For this example, use the write_dynamic_frame.from_options method.

```
# Example: Write CSV to S3
# For show, customize how we write string type values.  Set quoteChar to -1 so our
 values are not quoted.

from pyspark.context import SparkContext
from awsglue.context import GlueContext

sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)

glueContext.write_dynamic_frame.from_options(
    frame=dynamicFrame,
    connection_type="s3",
    connection_options={"path": "s3://s3path"},
    format="csv",
    format_options={
        "quoteChar": -1,
    },
)
```

You can also use DataFrames in a script (`pyspark.sql.DataFrame`).

```
dataFrame.write\
    .format("csv")\
    .option("quote", None)\
    .mode("append")\
    .save("s3://s3path")
```

Scala

For this example, use the getSinkWithFormat method.

```
// Example: Write CSV to S3
// For show, customize how we write string type values. Set quoteChar to -1 so our
 values are not quoted.

import com.amazonaws.services.glue.util.JsonOptions
import com.amazonaws.services.glue.{DynamicFrame, GlueContext}
import org.apache.spark.SparkContext

object GlueApp {
  def main(sysArgs: Array[String]): Unit = {
    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)

    glueContext.getSinkWithFormat(
        connectionType="s3",
        options=JsonOptions("""{"path": "s3://s3path"}"""),
        format="csv"
    ).writeDynamicFrame(dynamicFrame)
  }
}
```

You can also use DataFrames in a script (`org.apache.spark.sql.DataFrame`).

```
dataFrame.write
    .format("csv")
    .option("quote", null)
    .mode("Append")
    .save("s3://s3path")
```

CSV configuration reference

You can use the following `format_options` wherever AWS Glue libraries specify `format="csv"`:

- `separator` –Specifies the delimiter character. The default is a comma, but any other character can be specified.

  - **Type:** Text, **Default:** `","`
- `escaper` – Specifies a character to use for escaping. This option is used only when reading CSV files, not writing. If enabled, the character that immediately follows is used as-is, except for a small set of well-known escapes (`\n`, `\r`, `\t`, and `\0`).
  - **Type:** Text, **Default:** none
- `quoteChar` – Specifies the character to use for quoting. The default is a double quote. Set this to `-1` to turn off quoting entirely.
  - **Type:** Text, **Default:** `'"'`
- `multiLine` – Specifies whether a single record can span multiple lines. This can occur when a field contains a quoted new-line character. You must set this option to `True` if any record spans multiple lines. Enabling `multiLine` might decrease performance because it requires more cautious file-splitting while parsing.
  - **Type:** Boolean, **Default:** `false`
- `withHeader` – Specifies whether to treat the first line as a header. This option can be used in the `DynamicFrameReader` class.
  - **Type:** Boolean, **Default:** `false`
- `writeHeader` – Specifies whether to write the header to output. This option can be used in the `DynamicFrameWriter` class.
  - **Type:** Boolean, **Default:** `true`
- `skipFirst` – Specifies whether to skip the first data line.
  - **Type:** Boolean, **Default:** `false`
- `optimizePerformance` – Specifies whether to use the advanced SIMD CSV reader along with Apache Arrow–based columnar memory formats. Only available in AWS Glue 3.0+.
  - **Type:** Boolean, **Default:** `false`
- `strictCheckForQuoting` – When writing CSVs, Glue may add quotes to values it interprets as strings. This is done to prevent ambiguity in what is written out. To save time when deciding what to write, Glue may quote in certain situations where quotes are not necessary. Enabling a strict check will perform a more intensive computation and will only quote when strictly necessary. Only available in AWS Glue 3.0+.
  - **Type:** Boolean, **Default:** `false`

Optimize read performance with vectorized SIMD CSV reader

AWS Glue version 3.0 adds an optimized CSV reader that can significantly speed up overall job performance compared to row-based CSV readers.

The optimized reader:

- Uses CPU SIMD instructions to read from disk
- Immediately writes records to memory in a columnar format (Apache Arrow)
- Divides the records into batches

This saves processing time when records would be batched or converted to a columnar format later on. Some examples are when changing schemas or retrieving data by column.

To use the optimized reader, set `"optimizePerformance"` to `true` in the `format_options` or table property.

```
glueContext.create_dynamic_frame.from_options(
    frame = datasource1,
    connection_type = "s3",
    connection_options = {"paths": ["s3://s3path"]},
    format = "csv",
    format_options={
        "optimizePerformance": True,
        "separator": ","
        },
    transformation_ctx = "datasink2")
```

Limitations for the vectorized CSV reader

Note the following limitations of the vectorized CSV reader:

- It doesn't support the `multiLine` and `escaper` format options. It uses the default `escaper` of double quote char `'"'`. When these options are set, AWS Glue automatically falls back to using the row-based CSV reader.
- It doesn't support creating a DynamicFrame with ChoiceType.
- It doesn't support creating a DynamicFrame with error records.
- It doesn't support reading CSV files with multibyte characters such as Japanese or Chinese characters.

Using the Parquet format in AWS Glue

AWS Glue retrieves data from sources and writes data to targets stored and transported in various data formats. If your data is stored or transported in the Parquet data format, this document introduces you available features for using your data in AWS Glue.

AWS Glue supports using the Parquet format. This format is a performance-oriented, column-based data format. For an introduction to the format by the standard authority see, Apache Parquet Documentation Overview.

You can use AWS Glue to read Parquet files from Amazon S3 and from streaming sources as well as write Parquet files to Amazon S3. You can read and write `bzip` and `gzip` archives containing Parquet files from S3. You configure compression behavior on the S3 connection parameters instead of in the configuration discussed on this page.

The following table shows which common AWS Glue features support the Parquet format option.

Read	Write	Streaming read	Group small files	Job bookmarks
Supported	Supported	Supported	Unsupported	Supported*

* Supported in AWS Glue version 1.0+

Example: Read Parquet files or folders from S3

Prerequisites: You will need the S3 paths (`s3path`) to the Parquet files or folders that you want to read.

Configuration: In your function options, specify `format="parquet"`. In your `connection_options`, use the `paths` key to specify your `s3path`.

You can configure how the reader interacts with S3 in the `connection_options`. For details, see Connection types and options for ETL in AWS Glue: S3 connection parameters.

You can configure how the reader interprets Parquet files in your `format_options`. For details, see Parquet Configuration Reference.

The following AWS Glue ETL script shows the process of reading Parquet files or folders from S3:

Python

For this example, use the create_dynamic_frame.from_options method.

```
# Example: Read Parquet from S3

from pyspark.context import SparkContext
from awsglue.context import GlueContext

sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)
spark = glueContext.spark_session

dynamicFrame = glueContext.create_dynamic_frame.from_options(
    connection_type = "s3",
    connection_options = {"paths": ["s3://s3path/"]},
    format = "parquet"
)
```

You can also use DataFrames in a script (`pyspark.sql.DataFrame`).

```
dataFrame = spark.read.parquet("s3://s3path/")
```

Scala

For this example, use the getSourceWithFormat method.

```
// Example: Read Parquet from S3

import com.amazonaws.services.glue.util.JsonOptions
import com.amazonaws.services.glue.{DynamicFrame, GlueContext}
import org.apache.spark.SparkContext

object GlueApp {
  def main(sysArgs: Array[String]): Unit = {
    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)

    val dynamicFrame = glueContext.getSourceWithFormat(
      connectionType="s3",
      format="parquet",
      options=JsonOptions("""{"paths": ["s3://s3path"]}""")
    ).getDynamicFrame()
  }
}
```

You can also use DataFrames in a script (`org.apache.spark.sql.DataFrame`).

```
spark.read.parquet("s3://s3path/")
```

Example: Write Parquet files and folders to S3

Prerequisites: You will need an initialized DataFrame (`dataFrame`) or DynamicFrame (`dynamicFrame`). You will also need your expected S3 output path, `s3path`.

Configuration: In your function options, specify `format="parquet"`. In your `connection_options`, use the `paths` key to specify `s3path`.

You can further alter how the writer interacts with S3 in the `connection_options`. For details, see Connection types and options for ETL in AWS Glue: S3 connection parameters. You can configure how your operation writes the contents of your files in `format_options`. For details, see Parquet Configuration Reference.

The following AWS Glue ETL script shows the process of writing Parquet files and folders to S3.

We provide a custom Parquet writer with performance optimizations for DynamicFrames, through the `useGlueParquetWriter` configuration key. To determine if this writer is right for your workload, see Glue Parquet Writer.

Python

For this example, use the write_dynamic_frame.from_options method.

```
# Example: Write Parquet to S3
# Consider whether useGlueParquetWriter is right for your workflow.

from pyspark.context import SparkContext
from awsglue.context import GlueContext

sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)
```

```
glueContext.write_dynamic_frame.from_options(
    frame=dynamicFrame,
    connection_type="s3",
    format="parquet",
    connection_options={
        "path": "s3://s3path",
    },
    format_options={
        # "useGlueParquetWriter": True,
    },
)
```

You can also use DataFrames in a script (`pyspark.sql.DataFrame`).

```
df.write.parquet("s3://s3path/")
```

Scala

For this example, use the getSinkWithFormat method.

```
// Example: Write Parquet to S3
// Consider whether useGlueParquetWriter is right for your workflow.

import com.amazonaws.services.glue.util.JsonOptions
import com.amazonaws.services.glue.{DynamicFrame, GlueContext}
import org.apache.spark.SparkContext

object GlueApp {
  def main(sysArgs: Array[String]): Unit = {
    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)

    glueContext.getSinkWithFormat(
        connectionType="s3",
        options=JsonOptions("""{"path": "s3://s3path"}"""),
        format="parquet"
    ).writeDynamicFrame(dynamicFrame)
  }
}
```

You can also use DataFrames in a script (`org.apache.spark.sql.DataFrame`).

```
df.write.parquet("s3://s3path/")
```

Parquet configuration reference

You can use the following `format_options` wherever AWS Glue libraries specify `format="parquet"`:

- `useGlueParquetWriter` – Specifies the use of a custom Parquet writer that has performance optimizations for DynamicFrame workflows. For usage details, see Glue Parquet Writer.
  - **Type:** Boolean, **Default:**`false`
- `compression` – Specifies the compression codec used. Values are fully compatible with `org.apache.parquet.hadoop.metadata.CompressionCodecName`.
  - **Type:** Enumerated Text, **Default:** `"snappy"`
  - Values: `"uncompressed"`, `"snappy"`, `"gzip"`, and `"lzo"`
- `blockSize` – Specifies the size in bytes of a row group being buffered in memory. You use this for tuning performance. Size should divide exactly into a number of megabytes.
  - **Type:** Numerical, **Default:**`134217728`
  - The default value is equal to 128 MB.
- `pageSize` – Specifies the size in bytes of a page. You use this for tuning performance. A page is the smallest unit that must be read fully to access a single record.
  - **Type:** Numerical, **Default:**`1048576`
  - The default value is equal to 1 MB.

> **Note**
>
> Additionally, any options that are accepted by the underlying SparkSQL code can be passed to this format by way of the `connection_options` map parameter. For example, you can set a Spark configuration such as mergeSchema for the AWS Glue Spark reader to merge the schema for all files.

Optimize write performance with AWS Glue Parquet writer

Note

The AWS Glue Parquet writer has historically been accessed through the `glueparquet` format type. This access pattern is no longer advocated. Instead, use the `parquet` type with `useGlueParquetWriter` enabled.

The AWS Glue Parquet writer has performance enhancements that allow faster Parquet file writes. The traditional writer computes a schema before writing. The Parquet format doesn't store the schema in a quickly retrievable fashion, so this might take some time. With the AWS Glue Parquet writer, a pre-computed schema isn't required. The writer computes and modifies the schema dynamically, as data comes in.

Note the following limitations when you specify `useGlueParquetWriter`:

- The writer supports only schema evolution (such as adding or removing columns), but not changing column types, such as with `ResolveChoice`.
- The writer doesn't support writing empty DataFrames—for example, to write a schema-only file. When integrating with the AWS Glue Data Catalog by setting `enableUpdateCatalog=True`, attempting to write an empty DataFrame will not update the Data Catalog. This will result in creating a table in the Data Catalog without a schema.

If your transform doesn't require these limitations, turning on the AWS Glue Parquet writer should increase performance.

Using the XML format in AWS Glue

AWS Glue retrieves data from sources and writes data to targets stored and transported in various data formats. If your data is stored or transported in the XML data format, this document introduces you available features for using your data in AWS Glue.

AWS Glue supports using the XML format. This format represents highly configurable, rigidly defined data structures that aren't row or column based. XML is highly standardized. For an introduction to the format by the standard authority, see XML Essentials.

You can use AWS Glue to read XML files from Amazon S3, as well as `bzip` and `gzip` archives containing XML files. You configure compression behavior on the S3 connection parameters instead of in the configuration discussed on this page.

The following table shows which common AWS Glue features support the XML format option.

Read	Write	Streaming read	Group small files	Job bookmarks
Supported	Unsupported	Unsupported	Supported	Supported

Example: Read XML from S3

The XML reader takes an XML tag name. It examines elements with that tag within its input to infer a schema and populates a DynamicFrame with corresponding values. The AWS Glue XML functionality behaves similarly to the XML Data Source for Apache Spark. You might be able to gain insight around basic behavior by comparing this reader to that project's documentation.

Prerequisites: You will need the S3 paths (`s3path`) to the XML files or folders that you want to read, and some information about your XML file. You will also need the tag for the XML element you want to read, `xmlTag`.

Configuration: In your function options, specify `format="xml"`. In your `connection_options`, use the `paths` key to specify `s3path`. You can further configure how the reader interacts with S3 in the `connection_options`. For details, see Connection types and options for ETL in AWS Glue: S3 connection parameters. In your `format_options`, use the `rowTag` key to specify `xmlTag`. You can further configure how the reader interprets XML files in your `format_options`. For details, see XML Configuration Reference.

The following AWS Glue ETL script shows the process of reading XML files or folders from S3.

Python

For this example, use the create_dynamic_frame.from_options method.

```
# Example: Read XML from S3
# Set the rowTag option to configure the reader.

from awsglue.context import GlueContext
from pyspark.context import SparkContext

sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)

dynamicFrame = glueContext.create_dynamic_frame.from_options(
    connection_type="s3",
    connection_options={"paths": ["s3://s3path"]},
    format="xml",
    format_options={"rowTag": "xmlTag"},
)
```

You can also use DataFrames in a script (`pyspark.sql.DataFrame`).

```
dataFrame = spark.read\
    .format("xml")\
    .option("rowTag", "xmlTag")\
    .load("s3://s3path")
```

Scala

For this example, use the getSourceWithFormat operation.

```
// Example: Read XML from S3
// Set the rowTag option to configure the reader.

import com.amazonaws.services.glue.util.JsonOptions
import com.amazonaws.services.glue.GlueContext
import org.apache.spark.sql.SparkSession

val glueContext = new GlueContext(SparkContext.getOrCreate())
val sparkSession: SparkSession = glueContext.getSparkSession

object GlueApp {
  def main(sysArgs: Array[String]): Unit = {
    val dynamicFrame = glueContext.getSourceWithFormat(
      formatOptions=JsonOptions("""{"rowTag": "xmlTag"}"""),
      connectionType="s3",
      format="xml",
      options=JsonOptions("""{"paths": ["s3://s3path"], "recurse": true}""")
    ).getDynamicFrame()
}
```

You can also use DataFrames in a script (`org.apache.spark.sql.DataFrame`).

```
val dataFrame = spark.read
  .option("rowTag", "xmlTag")
  .format("xml")
  .load("s3://s3path“)
```

XML configuration reference

You can use the following `format_options` wherever AWS Glue libraries specify `format="xml"`:

- `rowTag` – Specifies the XML tag in the file to treat as a row. Row tags cannot be self-closing.
  - **Type:** Text, **Required**
- `encoding` – Specifies the character encoding. It can be the name or alias of a Charset supported by our runtime environment. We don't make specific guarantees around encoding support, but major encodings should work.
  - **Type:** Text, **Default:** `"UTF-8"`
- `excludeAttribute` – Specifies whether you want to exclude attributes in elements or not.
  - **Type:** Boolean, **Default:** `false`
- `treatEmptyValuesAsNulls` – Specifies whether to treat white space as a null value.
  - **Type:** Boolean, **Default:** `false`
- `attributePrefix` – A prefix for attributes to differentiate them from child element text. This prefix is used for field names.
  - **Type:** Text, **Default:** `"_"`
- `valueTag` – The tag used for a value when there are attributes in the element that have no child.
  - **Type:** Text, **Default:** `"_VALUE"`
- `ignoreSurroundingSpaces` – Specifies whether the white space that surrounds values should be ignored.
  - **Type:** Boolean, **Default:** `false`
- `withSchema` – Contains the expected schema, in situations where you want to override the inferred schema. If you don't use this option, AWS Glue infers the schema from the XML data.
  - **Type:** Text, **Default:** Not applicable

- The value should be a JSON object that represents a `StructType`.

Manually specify the XML schema

Manual XML schema example

This is an example of using the `withSchema` format option to specify the schema for XML data.

```
from awsglue.gluetypes import *

schema = StructType([
  Field("id", IntegerType()),
  Field("name", StringType()),
  Field("nested", StructType([
    Field("x", IntegerType()),
    Field("y", StringType()),
    Field("z", ChoiceType([IntegerType(), StringType()]))
  ]))
])

datasource0 = create_dynamic_frame_from_options(
    connection_type,
    connection_options={"paths": ["s3://xml_bucket/someprefix"]},
    format="xml",
    format_options={"withSchema": json.dumps(schema.jsonValue())},
    transformation_ctx = ""
)
```

Using the Avro format in AWS Glue

AWS Glue retrieves data from sources and writes data to targets stored and transported in various data formats. If your data is stored or transported in the Avro data format, this document introduces you available features for using your data in AWS Glue.

AWS Glue supports using the Avro format. This format is a performance-oriented, row-based data format. For an introduction to the format by the standard authority see, Apache Avro 1.8.2 Documentation.

You can use AWS Glue to read Avro files from Amazon S3 and from streaming sources as well as write Avro files to Amazon S3. You can read and write `bzip2` and `gzip` archives containing Avro files from S3. Additionally, you can write `deflate`, `snappy`, and `xz` archives containing

Avro files. You configure compression behavior on the S3 connection parameters instead of in the configuration discussed on this page.

The following table shows which common AWS Glue operations support the Avro format option.

Read	Write	Streaming read	Group small files	Job bookmarks
Supported	Supported	Supported*	Unsupported	Supported

*Supported with restrictions. For more information, see the section called "Notes and restrictions for Avro streaming sources".

Example: Read Avro files or folders from S3

Prerequisites: You will need the S3 paths (`s3path`) to the Avro files or folders that you want to read.

Configuration: In your function options, specify `format="avro"`. In your `connection_options`, use the `paths` key to specify `s3path`. You can configure how the reader interacts with S3 in the `connection_options`. For details, see Data format options for ETL inputs and outputs in AWS Glue: the section called "S3 connection parameters". You can configure how the reader interprets Avro files in your `format_options`. For details, see Avro Configuration Reference.

The following AWS Glue ETL script shows the process of reading Avro files or folders from S3:

Python

For this example, use the create_dynamic_frame.from_options method.

```
from pyspark.context import SparkContext
from awsglue.context import GlueContext

sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)

dynamicFrame = glueContext.create_dynamic_frame.from_options(
    connection_type="s3",
    connection_options={"paths": ["s3://s3path"]},
    format="avro"
)
```

Scala

For this example, use the getSourceWithFormat operation.

```
import com.amazonaws.services.glue.util.JsonOptions
import com.amazonaws.services.glue.GlueContext
import org.apache.spark.sql.SparkContext

object GlueApp {
  def main(sysArgs: Array[String]): Unit = {
    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)

    val dynamicFrame = glueContext.getSourceWithFormat(
      connectionType="s3",
      format="avro",
      options=JsonOptions("""{"paths": ["s3://s3path"]}""")
    ).getDynamicFrame()
  }
```

Example: Write Avro files and folders to S3

Prerequisites: You will need an initialized DataFrame (`dataFrame`) or DynamicFrame (`dynamicFrame`). You will also need your expected S3 output path, `s3path`.

Configuration: In your function options, specify `format="avro"`. In your `connection_options`, use the `paths` key to specify your `s3path`. You can further alter how the writer interacts with S3 in the `connection_options`. For details, see Data format options for ETL inputs and outputs in AWS Glue: the section called "S3 connection parameters". You can alter how the writer interprets Avro files in your `format_options`. For details, see Avro Configuration Reference.

The following AWS Glue ETL script shows the process of writing Avro files or folders to S3.

Python

For this example, use the write_dynamic_frame.from_options method.

```
from pyspark.context import SparkContext
from awsglue.context import GlueContext

sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)
```

```
glueContext.write_dynamic_frame.from_options(
    frame=dynamicFrame,
    connection_type="s3",
    format="avro",
    connection_options={
        "path": "s3://s3path"
    }
)
```

Scala

For this example, use the getSinkWithFormat method.

```
import com.amazonaws.services.glue.util.JsonOptions
import com.amazonaws.services.glue.{DynamicFrame, GlueContext}
import org.apache.spark.SparkContext

object GlueApp {
  def main(sysArgs: Array[String]): Unit = {
    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)

    glueContext.getSinkWithFormat(
      connectionType="s3",
      options=JsonOptions("""{"path": "s3://s3path"}"""),
      format="avro"
    ).writeDynamicFrame(dynamicFrame)
  }
}
```

Avro configuration reference

You can use the following `format_options` values wherever AWS Glue libraries specify `format="avro"`:

- `version` — Specifies the version of Apache Avro reader/writer format to support. The default is "1.7". You can specify `format_options={"version": “1.8”}` to enable Avro logical type reading and writing. For more information, see the Apache Avro 1.7.7 Specification and Apache Avro 1.8.2 Specification.

  The Apache Avro 1.8 connector supports the following logical type conversions:

For the reader: this table shows the conversion between Avro data type (logical type and Avro primitive type) and AWS Glue `DynamicFrame` data type for Avro reader 1.7 and 1.8.

Avro Data Type: Logical Type	Avro Data Type: Avro Primitive Type	GlueDynamicFrame Data Type: Avro Reader 1.7	GlueDynamicFrame Data Type: Avro Reader 1.8
Decimal	bytes	BINARY	Decimal
Decimal	fixed	BINARY	Decimal
Date	int	INT	Date
Time (millisecond)	int	INT	INT
Time (microsecond)	long	LONG	LONG
Timestamp (millisecond)	long	LONG	Timestamp
Timestamp (microsecond)	long	LONG	LONG
Duration (not a logical type)	fixed of 12	BINARY	BINARY

For the writer: this table shows the conversion between AWS Glue `DynamicFrame` data type and Avro data type for Avro writer 1.7 and 1.8.

AWS Glue `DynamicFrame` Data Type	Avro Data Type: Avro Writer 1.7	Avro Data Type: Avro Writer 1.8
Decimal	String	decimal
Date	String	date
Timestamp	String	timestamp-micros

Avro Spark DataFrame support

In order to use Avro from the Spark DataFrame API, you need to install the Spark Avro plugin for the corresponding Spark version. The version of Spark available in your job is determined by your AWS Glue version. For more information about Spark versions, see the section called "AWS Glue versions". This plugin is maintained by Apache, we do not make specific guarantees of support.

In AWS Glue 2.0 - use version 2.4.3 of the Spark Avro plugin. You can find this JAR on Maven Central, see org.apache.spark:spark-avro_2.12:2.4.3.

In AWS Glue 3.0 - use version 3.1.1 of the Spark Avro plugin. You can find this JAR on Maven Central, see org.apache.spark:spark-avro_2.12:3.1.1.

To include extra JARs in a AWS Glue ETL job, use the `--extra-jars` job parameter. For more information about job parameters, see the section called "Job parameters". You can also configure this parameter in the AWS Management Console.

Using the grokLog format in AWS Glue

AWS Glue retrieves data from sources and writes data to targets stored and transported in various data formats. If your data is stored or transported in a loosely structured plaintext format, this document introduces you available features for using your data in AWS Glue through Grok patterns.

AWS Glue supports using Grok patterns. Grok patterns are similar to regular expression capture groups. They recognize patterns of character sequences in a plaintext file and give them a type and purpose. In AWS Glue, their primary purpose is to read logs. For an introduction to the Grok by the authors, see Logstash Reference: Grok filter plugin.

Read	Write	Streaming read	Group small files	Job bookmarks
Supported	Not Applicable	Supported	Supported	Unsupported

grokLog configuration reference

You can use the following `format_options` values with `format="grokLog"`:

- `logFormat` — Specifies the Grok pattern that matches the log's format.

- `customPatterns` — Specifies additional Grok patterns used here.
- `MISSING` — Specifies the signal to use in identifying missing values. The default is `'-'`.
- `LineCount` — Specifies the number of lines in each log record. The default is `'1'`, and currently only single-line records are supported.
- `StrictMode` — A Boolean value that specifies whether strict mode is turned on. In strict mode, the reader doesn't do automatic type conversion or recovery. The default value is `"false"`.

Using the Ion format in AWS Glue

AWS Glue retrieves data from sources and writes data to targets stored and transported in various data formats. If your data is stored or transported in the Ion data format, this document introduces you available features for using your data in AWS Glue.

AWS Glue supports using the Ion format. This format represents data structures (that aren't row or column based) in interchangeable binary and plaintext representations. For an introduction to the format by the authors, see Amazon Ion. (For more information, see the Amazon Ion Specification.)

You can use AWS Glue to read Ion files from Amazon S3. You can read `bzip` and `gzip` archives containing Ion files from S3. You configure compression behavior on the S3 connection parameters instead of in the configuration discussed on this page.

The following table shows which common AWS Glue operations support the Ion format option.

Read	Write	Streaming read	Group small files	Job bookmarks
Supported	Unsupported	Unsupported	Supported	Unsupported

Example: Read Ion files and folders from S3

Prerequisites: You will need the S3 paths (`s3path`) to the Ion files or folders that you want to read.

Configuration: In your function options, specify `format="json"`. In your `connection_options`, use the `paths` key to specify your `s3path`. You can configure how the reader interacts with S3 in the `connection_options`. For details, see Connection types and options for ETL in AWS Glue: the section called “S3 connection parameters”.

The following AWS Glue ETL script shows the process of reading Ion files or folders from S3:

Python

For this example, use the create_dynamic_frame.from_options method.

```
# Example: Read ION from S3

from pyspark.context import SparkContext
from awsglue.context import GlueContext

sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)

dynamicFrame = glueContext.create_dynamic_frame.from_options(
    connection_type="s3",
    connection_options={"paths": ["s3://s3path"]},
    format="ion"
)
```

Scala

For this example, use the getSourceWithFormat operation.

```
// Example: Read ION from S3

import com.amazonaws.services.glue.util.JsonOptions
import com.amazonaws.services.glue.GlueContext
import org.apache.spark.SparkContext

object GlueApp {
  def main(sysArgs: Array[String]): Unit = {
    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)

    val dynamicFrame = glueContext.getSourceWithFormat(
      connectionType="s3",
      format="ion",
      options=JsonOptions("""{"paths": ["s3://s3path"], "recurse": true}""")
    ).getDynamicFrame()
  }
}
```

Ion configuration reference

There are no `format_options` values for `format="ion"`.

Using the JSON format in AWS Glue

AWS Glue retrieves data from sources and writes data to targets stored and transported in various data formats. If your data is stored or transported in the JSON data format, this document introduces you to available features for using your data in AWS Glue.

AWS Glue supports using the JSON format. This format represents data structures with consistent shape but flexible contents, that aren't row or column based. JSON is defined by parallel standards issued by several authorities, one of which is ECMA-404. For an introduction to the format by a commonly referenced source, see Introducing JSON.

You can use AWS Glue to read JSON files from Amazon S3, as well as `bzip` and `gzip` compressed JSON files. You configure compression behavior on the S3 connection parameters instead of in the configuration discussed on this page.

Read	Write	Streaming read	Group small files	Job bookmarks
Supported	Supported	Supported	Supported	Supported

Example: Read JSON files or folders from S3

Prerequisites: You will need the S3 paths (`s3path`) to the JSON files or folders you would like to read.

Configuration: In your function options, specify `format="json"`. In your `connection_options`, use the `paths` key to specify your `s3path`. You can further alter how your read operation will traverse s3 in the connection options, consult the section called "S3 connection parameters" for details. You can configure how the reader interprets JSON files in your `format_options`. For details, see JSON Configuration Reference.

The following AWS Glue ETL script shows the process of reading JSON files or folders from S3:

Python

For this example, use the create_dynamic_frame.from_options method.

```
# Example: Read JSON from S3
# For show, we handle a nested JSON file that we can limit with the JsonPath
 parameter
# For show, we also handle a JSON where a single entry spans multiple lines
# Consider whether optimizePerformance is right for your workflow.

from pyspark.context import SparkContext
from awsglue.context import GlueContext

sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)
spark = glueContext.spark_session

dynamicFrame = glueContext.create_dynamic_frame.from_options(
    connection_type="s3",
    connection_options={"paths": ["s3://s3path"]},
    format="json",
    format_options={
        "jsonPath": "$.id",
        "multiline": True,
        # "optimizePerformance": True, -> not compatible with jsonPath, multiline
    }
)
```

You can also use DataFrames in a script (`pyspark.sql.DataFrame`).

```
dataFrame = spark.read\
    .option("multiline", "true")\
    .json("s3://s3path")
```

Scala

For this example, use the getSourceWithFormat operation.

```
// Example: Read JSON from S3
// For show, we handle a nested JSON file that we can limit with the JsonPath
 parameter
// For show, we also handle a JSON where a single entry spans multiple lines
// Consider whether optimizePerformance is right for your workflow.

import com.amazonaws.services.glue.util.JsonOptions
import com.amazonaws.services.glue.{DynamicFrame, GlueContext}
import org.apache.spark.SparkContext

object GlueApp {
  def main(sysArgs: Array[String]): Unit = {
    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)

    val dynamicFrame = glueContext.getSourceWithFormat(
      formatOptions=JsonOptions("""{"jsonPath": "$.id", "multiline": true,
 "optimizePerformance":false}"""),
      connectionType="s3",
      format="json",
      options=JsonOptions("""{"paths": ["s3://s3path"], "recurse": true}""")
    ).getDynamicFrame()
  }
}
```

You can also use DataFrames in a script (`pyspark.sql.DataFrame`).

```
val dataFrame = spark.read
    .option("multiline", "true")
    .json("s3://s3path")
```

Example: Write JSON files and folders to S3

Prerequisites: You will need an initialized DataFrame (`dataFrame`) or DynamicFrame (`dynamicFrame`). You will also need your expected S3 output path, `s3path`.

Configuration: In your function options, specify `format="json"`. In your `connection_options`, use the `paths` key to specify `s3path`. You can further alter how the writer interacts with S3 in the `connection_options`. For details, see Data format options for ETL inputs and outputs in AWS Glue : the section called “S3 connection parameters”. You can configure how the writer interprets JSON files in your `format_options`. For details, see JSON Configuration Reference.

The following AWS Glue ETL script shows the process of writing JSON files or folders from S3:

Python

For this example, use the write_dynamic_frame.from_options method.

```
# Example: Write JSON to S3
```

```
from pyspark.context import SparkContext
from awsglue.context import GlueContext

sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)

glueContext.write_dynamic_frame.from_options(
    frame=dynamicFrame,
    connection_type="s3",
    connection_options={"path": "s3://s3path"},
    format="json"
)
```

You can also use DataFrames in a script (`pyspark.sql.DataFrame`).

```
df.write.json("s3://s3path/")
```

Scala

For this example, use the getSinkWithFormat method.

```
// Example: Write JSON to S3

import com.amazonaws.services.glue.util.JsonOptions
import com.amazonaws.services.glue.{DynamicFrame, GlueContext}
import org.apache.spark.SparkContext

object GlueApp {
  def main(sysArgs: Array[String]): Unit = {
    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)

    glueContext.getSinkWithFormat(
        connectionType="s3",
        options=JsonOptions("""{"path": "s3://s3path"}"""),
        format="json"
    ).writeDynamicFrame(dynamicFrame)
  }
}
```

You can also use DataFrames in a script (`pyspark.sql.DataFrame`).

```
df.write.json("s3://s3path")
```

Json configuration reference

You can use the following `format_options` values with `format="json"`:

- `jsonPath` — A JsonPath expression that identifies an object to be read into records. This is particularly useful when a file contains records nested inside an outer array. For example, the following JsonPath expression targets the `id` field of a JSON object.

  ```
  format="json", format_options={"jsonPath": "$.id"}
  ```
- `multiline` — A Boolean value that specifies whether a single record can span multiple lines. This can occur when a field contains a quoted new-line character. You must set this option to `"true"` if any record spans multiple lines. The default value is `"false"`, which allows for more aggressive file-splitting during parsing.
- `optimizePerformance` — A Boolean value that specifies whether to use the advanced SIMD JSON reader along with Apache Arrow based columnar memory formats. Only available in AWS Glue 3.0. Not compatible with `multiline` or `jsonPath`. Providing either of those options will instruct AWS Glue to fall back to the standard reader.
- `withSchema` — A String value that specifies a table schema in the format described in the section called “Specify XML schema”. Only used with `optimizePerformance` when reading from non-Catalog connections.

Using vectorized SIMD JSON reader with Apache Arrow columnar format

AWS Glue version 3.0 adds a vectorized reader for JSON data. It performs 2x faster under certain conditions, compared to the standard reader. This reader comes with certain limitations users should be aware of before use, documented in this section.

To use the optimized reader, set `"optimizePerformance"` to True in the `format_options` or table property. You will also need to provide `withSchema` unless reading from the catalog. `withSchema` expects an input as described in the the section called “Specify XML schema”

```

// Read from S3 data source
glueContext.create_dynamic_frame.from_options(
```

```
    connection_type = "s3",
    connection_options = {"paths": ["s3://s3path"]},
    format = "json",
    format_options={
        "optimizePerformance": True,
        "withSchema": SchemaString
        })

// Read from catalog table
glueContext.create_dynamic_frame.from_catalog(
    database = database,
    table_name = table,
    additional_options = {
    // The vectorized reader for JSON can read your schema from a catalog table
 property.
        "optimizePerformance": True,
        })
```

For more information about the building a *`SchemaString`* in the AWS Glue library, see the section called "Types".

Limitations for the vectorized CSV reader

Note the following limitations:

- JSON elements with nested objects or array values are not supported. If provided, AWS Glue will fall back to the standard reader.
- A schema must be provided, either from the Catalog or with `withSchema`.
- Not compatible with `multiline` or `jsonPath`. Providing either of those options will instruct AWS Glue to fall back to the standard reader.
- Providing input records that do not match the input schema will cause the reader to fail.
- Error records will not be created.
- JSON files with multi-byte characters (such as Japanese or Chinese characters) are not supported.

Using the ORC format in AWS Glue

AWS Glue retrieves data from sources and writes data to targets stored and transported in various data formats. If your data is stored or transported in the ORC data format, this document introduces you available features for using your data in AWS Glue.

AWS Glue supports using the ORC format. This format is a performance-oriented, column-based data format. For an introduction to the format by the standard authority see, Apache Orc.

You can use AWS Glue to read ORC files from Amazon S3 and from streaming sources as well as write ORC files to Amazon S3. You can read and write `bzip` and `gzip` archives containing ORC files from S3. You configure compression behavior on the S3 connection parameters instead of in the configuration discussed on this page.

The following table shows which common AWS Glue operations support the ORC format option.

Read	Write	Streaming read	Group small files	Job bookmarks
Supported	Supported	Supported	Unsupported	Supported*

*Supported in AWS Glue version 1.0+

Example: Read ORC files or folders from S3

Prerequisites: You will need the S3 paths (`s3path`) to the ORC files or folders that you want to read.

Configuration: In your function options, specify `format="orc"`. In your `connection_options`, use the `paths` key to specify your `s3path`. You can configure how the reader interacts with S3 in the `connection_options`. For details, see Connection types and options for ETL in AWS Glue: the section called "S3 connection parameters".

The following AWS Glue ETL script shows the process of reading ORC files or folders from S3:

Python

For this example, use the create_dynamic_frame.from_options method.

```
from pyspark.context import SparkContext
from awsglue.context import GlueContext

sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)

dynamicFrame = glueContext.create_dynamic_frame.from_options(
    connection_type="s3",
    connection_options={"paths": ["s3://s3path"]},
    format="orc"
)
```

You can also use DataFrames in a script (`pyspark.sql.DataFrame`).

```
dataFrame = spark.read\
    .orc("s3://s3path")
```

Scala

For this example, use the getSourceWithFormat operation.

```
import com.amazonaws.services.glue.util.JsonOptions
import com.amazonaws.services.glue.GlueContext
import org.apache.spark.sql.SparkContext

object GlueApp {
  def main(sysArgs: Array[String]): Unit = {
    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)

    val dynamicFrame = glueContext.getSourceWithFormat(
      connectionType="s3",
      format="orc",
      options=JsonOptions("""{"paths": ["s3://s3path"]}""")
    ).getDynamicFrame()
  }
}
```

You can also use DataFrames in a script (`pyspark.sql.DataFrame`).

```
val dataFrame = spark.read
    .orc("s3://s3path")
```

Example: Write ORC files and folders to S3

Prerequisites: You will need an initialized DataFrame (`dataFrame`) or DynamicFrame (`dynamicFrame`). You will also need your expected S3 output path, `s3path`.

Configuration: In your function options, specify `format="orc"`. In your connection options, use the `paths` key to specify `s3path`. You can further alter how the writer interacts with S3 in the `connection_options`. For details, see Data format options for ETL inputs and outputs in AWS Glue: the section called “S3 connection parameters”. The following code example shows the process:

Python

For this example, use the write_dynamic_frame.from_options method.

```
from pyspark.context import SparkContext
from awsglue.context import GlueContext

sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)

glueContext.write_dynamic_frame.from_options(
    frame=dynamicFrame,
    connection_type="s3",
    format="orc",
    connection_options={
        "path": "s3://s3path"
    }
)
```

You can also use DataFrames in a script (`pyspark.sql.DataFrame`).

```
df.write.orc("s3://s3path/")
```

Scala

For this example, use the getSinkWithFormat method.

```
import com.amazonaws.services.glue.util.JsonOptions
import com.amazonaws.services.glue.{DynamicFrame, GlueContext}
import org.apache.spark.SparkContext

object GlueApp {
  def main(sysArgs: Array[String]): Unit = {
    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)
    glueContext.getSinkWithFormat(
      connectionType="s3",
      options=JsonOptions("""{"path": "s3://s3path"}"""),
      format="orc"
    ).writeDynamicFrame(dynamicFrame)
  }
}
```

You can also use DataFrames in a script (`pyspark.sql.DataFrame`).

```
df.write.orc("s3://s3path/")
```

ORC configuration reference

There are no `format_options` values for `format="orc"`. However, any options that are accepted by the underlying SparkSQL code can be passed to it by way of the `connection_options` map parameter.

Using data lake frameworks with AWS Glue ETL jobs

Open-source data lake frameworks simplify incremental data processing for files that you store in data lakes built on Amazon S3. AWS Glue 3.0 and later supports the following open-source data lake frameworks:

- Apache Hudi
- Linux Foundation Delta Lake
- Apache Iceberg

We provide native support for these frameworks so that you can read and write data that you store in Amazon S3 in a transactionally consistent manner. There's no need to install a separate connector or complete extra configuration steps in order to use these frameworks in AWS Glue ETL jobs.

When you manage datasets through the AWS Glue Data Catalog, you can use AWS Glue methods to read and write data lake tables with Spark DataFrames. You can also read and write Amazon S3 data using the Spark DataFrame API.

In this video, you can learn about the basics of how Apache Hudi, Apache Iceberg, and Delta Lake work. You'll see how to insert, update, and delete data in your data lake and how each of these frameworks works.

Topics

- Limitations
- Using the Hudi framework in AWS Glue
- Using the Delta Lake framework in AWS Glue
- Using the Iceberg framework in AWS Glue

Limitations

Consider the following limitations before you use data lake frameworks with AWS Glue.

- The following AWS Glue `GlueContext` methods for DynamicFrame don't support reading and writing data lake framework tables. Use the `GlueContext` methods for DataFrame or Spark DataFrame API instead.
  - `create_dynamic_frame.from_catalog`
  - `write_dynamic_frame.from_catalog`
  - `getDynamicFrame`
  - `writeDynamicFrame`
- The following `GlueContext` methods for DataFrame are supported with Lake Formation permission control:
  - `create_data_frame.from_catalog`
  - `write_data_frame.from_catalog`
  - `getDataFrame`
  - `writeDataFrame`
- Grouping small files is not supported.
- Job bookmarks are not supported.
- Apache Hudi 0.10.1 for AWS Glue 3.0 doesn't support Hudi Merge on Read (MoR) tables.
- `ALTER TABLE … RENAME TO` is not available for Apache Iceberg 0.13.1 for AWS Glue 3.0.

Limitations for data lake format tables managed by Lake Formation permissions

The data lake formats are integrated with AWS Glue ETL via Lake Formation permissions. Creating a DynamicFrame using `create_dynamic_frame` is not supported. For more information, see the following examples:

- Example: Read and write Iceberg table with Lake Formation permission control
- Example: Read and write Hudi table with Lake Formation permission control
- Example: Read and write Delta Lake table with Lake Formation permission control

> **Note**
>
> The integration with AWS Glue ETL via Lake Formation permissions for Apache Hudi, Apache Iceberg, and Delta Lake is supported only in AWS Glue version 4.0.

Apache Iceberg has the best integration with AWS Glue ETL via Lake Formation permissions. It supports almost all operations and includes SQL support.

Hudi supports most basic operations with the exception of administrative operations. This is because these options generally are done via writing of dataframes and specified via `additional_options`. You need to use AWS Glue APIs to create DataFrames for your operations as SparkSQL is not supported.

Delta Lake only supports the reading and appending and overwriting of table data. Delta Lake requires the use of their own libraries to be able to perform various tasks such as updates.

The following features are not available for Iceberg tables managed by Lake Formation permissions.

- Compaction using AWS Glue ETL
- Spark SQL support via AWS Glue ETL

The following are limitations of Hudi tables managed by Lake Formation permissions:

- Removal of orphaned files

The following are limitations of Delta Lake tables managed by Lake Formation permissions:

- All features other than inserting and reading from Delta Lake tables.

Using the Hudi framework in AWS Glue

AWS Glue 3.0 and later supports Apache Hudi framework for data lakes. Hudi is an open-source data lake storage framework that simplifies incremental data processing and data pipeline development. This topic covers available features for using your data in AWS Glue when you transport or store your data in a Hudi table. To learn more about Hudi, see the official Apache Hudi documentation.

You can use AWS Glue to perform read and write operations on Hudi tables in Amazon S3, or work with Hudi tables using the AWS Glue Data Catalog. Additional operations including insert, update, and all of the Apache Spark operations are also supported.

> **Note**
>
> The Apache Hudi 0.15.0 implementation in AWS Glue 5.0 internally reverts HUDI-7001. It does not exhibit the regression related to Complex Key gen when the record key consists of a single field. However this behavior differs from OSS Apache Hudi 0.15.0.
> Apache Hudi 0.10.1 for AWS Glue 3.0 doesn't support Hudi Merge on Read (MoR) tables.

The following table lists the Hudi version that is included in each AWS Glue version.

AWS Glue version	Supported Hudi version
5.0	0.15.0
4.0	0.12.1
3.0	0.10.1

To learn more about the data lake frameworks that AWS Glue supports, see Using data lake frameworks with AWS Glue ETL jobs.

Enabling Hudi

To enable Hudi for AWS Glue, complete the following tasks:

- Specify `hudi` as a value for the `--datalake-formats` job parameter. For more information, see Using job parameters in AWS Glue jobs.
- Create a key named `--conf` for your AWS Glue job, and set it to the following value. Alternatively, you can set the following configuration using `SparkConf` in your script. These settings help Apache Spark correctly handle Hudi tables.

  ```
  spark.serializer=org.apache.spark.serializer.KryoSerializer
  ```
- Lake Formation permission support for Hudi is enabled by default for AWS Glue 4.0. No additional configuration is needed for reading/writing to Lake Formation-registered Hudi tables. To read a registered Hudi table, the AWS Glue job IAM role must have the SELECT permission. To write to a registered Hudi table, the AWS Glue job IAM role must have the SUPER permission. To learn more about managing Lake Formation permissions, see Granting and revoking permissions on Data Catalog resources.

Using a different Hudi version

To use a version of Hudi that AWS Glue doesn't support, specify your own Hudi JAR files using the `--extra-jars` job parameter. Do not include `hudi` as a value for the `--datalake-formats` job parameter. If you use AWS Glue 5.0, you must set `--user-jars-first true` job parameter.

Example: Write a Hudi table to Amazon S3 and register it in the AWS Glue Data Catalog

This example script demonstrates how to write a Hudi table to Amazon S3 and register the table to the AWS Glue Data Catalog. The example uses the Hudi Hive Sync tool to register the table.

> **Note**
>
> This example requires you to set the `--enable-glue-datacatalog` job parameter in order to use the AWS Glue Data Catalog as an Apache Spark Hive metastore. To learn more, see Using job parameters in AWS Glue jobs.

Python

```
# Example: Create a Hudi table from a DataFrame
# and register the table to Glue Data Catalog

```

```
additional_options={
    "hoodie.table.name": "<your_table_name>",
    "hoodie.database.name": "<your_database_name>",
    "hoodie.datasource.write.storage.type": "COPY_ON_WRITE",
    "hoodie.datasource.write.operation": "upsert",
    "hoodie.datasource.write.recordkey.field": "<your_recordkey_field>",
    "hoodie.datasource.write.precombine.field": "<your_precombine_field>",
    "hoodie.datasource.write.partitionpath.field": "<your_partitionkey_field>",
    "hoodie.datasource.write.hive_style_partitioning": "true",
    "hoodie.datasource.hive_sync.enable": "true",
    "hoodie.datasource.hive_sync.database": "<your_database_name>",
    "hoodie.datasource.hive_sync.table": "<your_table_name>",
    "hoodie.datasource.hive_sync.partition_fields": "<your_partitionkey_field>",
    "hoodie.datasource.hive_sync.partition_extractor_class":
 "org.apache.hudi.hive.MultiPartKeysValueExtractor",
    "hoodie.datasource.hive_sync.use_jdbc": "false",
    "hoodie.datasource.hive_sync.mode": "hms",
    "path": "s3://<s3Path/>"
}

dataFrame.write.format("hudi") \
    .options(**additional_options) \
    .mode("overwrite") \
    .save()
```

Scala

```
// Example: Example: Create a Hudi table from a DataFrame
// and register the table to Glue Data Catalog

val additionalOptions = Map(
  "hoodie.table.name" -> "<your_table_name>",
  "hoodie.database.name" -> "<your_database_name>",
  "hoodie.datasource.write.storage.type" -> "COPY_ON_WRITE",
  "hoodie.datasource.write.operation" -> "upsert",
  "hoodie.datasource.write.recordkey.field" -> "<your_recordkey_field>",
  "hoodie.datasource.write.precombine.field" -> "<your_precombine_field>",
  "hoodie.datasource.write.partitionpath.field" -> "<your_partitionkey_field>",
  "hoodie.datasource.write.hive_style_partitioning" -> "true",
  "hoodie.datasource.hive_sync.enable" -> "true",
  "hoodie.datasource.hive_sync.database" -> "<your_database_name>",
  "hoodie.datasource.hive_sync.table" -> "<your_table_name>",
  "hoodie.datasource.hive_sync.partition_fields" -> "<your_partitionkey_field>",
  "hoodie.datasource.hive_sync.partition_extractor_class" ->
 "org.apache.hudi.hive.MultiPartKeysValueExtractor",
  "hoodie.datasource.hive_sync.use_jdbc" -> "false",
  "hoodie.datasource.hive_sync.mode" -> "hms",
  "path" -> "s3://<s3Path/>")

dataFrame.write.format("hudi")
  .options(additionalOptions)
  .mode("append")
  .save()
```

Example: Read a Hudi table from Amazon S3 using the AWS Glue Data Catalog

This example reads the Hudi table that you created in the Example: Write a Hudi table to Amazon S3 and register it in the AWS Glue Data Catalog from Amazon S3.

Note

This example requires you to set the `--enable-glue-datacatalog` job parameter in order to use the AWS Glue Data Catalog as an Apache Spark Hive metastore. To learn more, see Using job parameters in AWS Glue jobs.

Python

For this example, use the `GlueContext.create_data_frame.from_catalog()` method.

```
# Example: Read a Hudi table from Glue Data Catalog

from awsglue.context import GlueContext
from pyspark.context import SparkContext

sc = SparkContext()
glueContext = GlueContext(sc)

dataFrame = glueContext.create_data_frame.from_catalog(
    database = "<your_database_name>",
    table_name = "<your_table_name>"
)
```

Scala

For this example, use the getCatalogSource method.

```
// Example: Read a Hudi table from Glue Data Catalog

import com.amazonaws.services.glue.GlueContext
import org.apache.spark.SparkContext

object GlueApp {
  def main(sysArgs: Array[String]): Unit = {
    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)

    val dataFrame = glueContext.getCatalogSource(
      database = "<your_database_name>",
      tableName = "<your_table_name>"
    ).getDataFrame()
  }
}
```

Example: Update and insert a `DataFrame` into a Hudi table in Amazon S3

This example uses the AWS Glue Data Catalog to insert a DataFrame into the Hudi table that you created in Example: Write a Hudi table to Amazon S3 and register it in the AWS Glue Data Catalog.

> **Note**
>
> This example requires you to set the `--enable-glue-datacatalog` job parameter in order to use the AWS Glue Data Catalog as an Apache Spark Hive metastore. To learn more, see Using job parameters in AWS Glue jobs.

Python

For this example, use the `GlueContext.write_data_frame.from_catalog()` method.

```
# Example: Upsert a Hudi table from Glue Data Catalog

from awsglue.context import GlueContext
from pyspark.context import SparkContext

sc = SparkContext()
glueContext = GlueContext(sc)

glueContext.write_data_frame.from_catalog(
    frame = dataFrame,
    database = "<your_database_name>",
    table_name = "<your_table_name>",
    additional_options={
        "hoodie.table.name": "<your_table_name>",
        "hoodie.database.name": "<your_database_name>",
        "hoodie.datasource.write.storage.type": "COPY_ON_WRITE",
        "hoodie.datasource.write.operation": "upsert",
        "hoodie.datasource.write.recordkey.field": "<your_recordkey_field>",
        "hoodie.datasource.write.precombine.field": "<your_precombine_field>",
        "hoodie.datasource.write.partitionpath.field": "<your_partitionkey_field>",
        "hoodie.datasource.write.hive_style_partitioning": "true",
        "hoodie.datasource.hive_sync.enable": "true",
        "hoodie.datasource.hive_sync.database": "<your_database_name>",
        "hoodie.datasource.hive_sync.table": "<your_table_name>",
        "hoodie.datasource.hive_sync.partition_fields": "<your_partitionkey_field>",
        "hoodie.datasource.hive_sync.partition_extractor_class":
 "org.apache.hudi.hive.MultiPartKeysValueExtractor",
        "hoodie.datasource.hive_sync.use_jdbc": "false",
        "hoodie.datasource.hive_sync.mode": "hms"
    }
)
```

Scala

For this example, use the getCatalogSink method.

```
// Example: Upsert a Hudi table from Glue Data Catalog

import com.amazonaws.services.glue.GlueContext
import com.amazonaws.services.glue.util.JsonOptions
import org.apacke.spark.SparkContext

object GlueApp {
  def main(sysArgs: Array[String]): Unit = {
    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)
    glueContext.getCatalogSink("<your_database_name>", "<your_table_name>",
      additionalOptions = JsonOptions(Map(
        "hoodie.table.name" -> "<your_table_name>",
        "hoodie.database.name" -> "<your_database_name>",
        "hoodie.datasource.write.storage.type" -> "COPY_ON_WRITE",
        "hoodie.datasource.write.operation" -> "upsert",
        "hoodie.datasource.write.recordkey.field" -> "<your_recordkey_field>",
        "hoodie.datasource.write.precombine.field" -> "<your_precombine_field>",
        "hoodie.datasource.write.partitionpath.field" ->
 "<your_partitionkey_field>",
        "hoodie.datasource.write.hive_style_partitioning" -> "true",
        "hoodie.datasource.hive_sync.enable" -> "true",
        "hoodie.datasource.hive_sync.database" -> "<your_database_name>",
        "hoodie.datasource.hive_sync.table" -> "<your_table_name>",
        "hoodie.datasource.hive_sync.partition_fields" ->
 "<your_partitionkey_field>",
        "hoodie.datasource.hive_sync.partition_extractor_class" ->
 "org.apache.hudi.hive.MultiPartKeysValueExtractor",
        "hoodie.datasource.hive_sync.use_jdbc" -> "false",
        "hoodie.datasource.hive_sync.mode" -> "hms"
      )))
      .writeDataFrame(dataFrame, glueContext)
  }
}
```

Example: Read a Hudi table from Amazon S3 using Spark

This example reads a Hudi table from Amazon S3 using the Spark DataFrame API.

Python

```
# Example: Read a Hudi table from S3 using a Spark DataFrame

dataFrame = spark.read.format("hudi").load("s3://<s3path/>")
```

Scala

```
// Example: Read a Hudi table from S3 using a Spark DataFrame

val dataFrame = spark.read.format("hudi").load("s3://<s3path/>")
```

Example: Write a Hudi table to Amazon S3 using Spark

This example writes a Hudi table to Amazon S3 using Spark.

Python

```
# Example: Write a Hudi table to S3 using a Spark DataFrame

dataFrame.write.format("hudi") \
    .options(**additional_options) \
    .mode("overwrite") \
    .save("s3://<s3Path/>)
```

Scala

```
// Example: Write a Hudi table to S3 using a Spark DataFrame

dataFrame.write.format("hudi")
  .options(additionalOptions)
  .mode("overwrite")
  .save("s3://<s3path/>")
```

Example: Read and write Hudi table with Lake Formation permission control

This example reads and writes a Hudi table with Lake Formation permission control.

1. Create a Hudi table and register it in Lake Formation.
   a. To enable Lake Formation permission control, you'll first need to register the table Amazon S3 path on Lake Formation. For more information, see Registering an Amazon S3 location. You can register it either from the Lake Formation console or by using the AWS CLI:

   ```
   aws lakeformation register-resource --resource-arn arn:aws:s3:::<s3-bucket>/<s3-
   folder> --use-service-linked-role --region <REGION>
   ```

   Once you register an Amazon S3 location, any AWS Glue table pointing to the location (or any of its child locations) will return the value for the `IsRegisteredWithLakeFormation` parameter as true in the `GetTable` call.

   b. Create a Hudi table that points to the registered Amazon S3 path through the Spark dataframe API:

```
hudi_options = {
    'hoodie.table.name': table_name,
    'hoodie.database.name': database_name,
    'hoodie.datasource.write.storage.type': 'COPY_ON_WRITE',
    'hoodie.datasource.write.recordkey.field': 'product_id',
    'hoodie.datasource.write.table.name': table_name,
    'hoodie.datasource.write.operation': 'upsert',
    'hoodie.datasource.write.precombine.field': 'updated_at',
    'hoodie.datasource.write.hive_style_partitioning': 'true',
    'hoodie.upsert.shuffle.parallelism': 2,
    'hoodie.insert.shuffle.parallelism': 2,
    'path': <S3_TABLE_LOCATION>,
    'hoodie.datasource.hive_sync.enable': 'true',
    'hoodie.datasource.hive_sync.database': database_name,
    'hoodie.datasource.hive_sync.table': table_name,
    'hoodie.datasource.hive_sync.use_jdbc': 'false',
    'hoodie.datasource.hive_sync.mode': 'hms'
}

df_products.write.format("hudi")  \
    .options(**hudi_options)  \
    .mode("overwrite")  \
    .save()
```

2. Grant Lake Formation permission to the AWS Glue job IAM role. You can either grant permissions from the Lake Formation console, or using the AWS CLI. For more information, see Granting table permissions using the Lake Formation console and the named resource method

3. Read the Hudi table registered in Lake Formation. The code is same as reading a non-registered Hudi table. Note that the AWS Glue job IAM role needs to have the SELECT permission for the read to succeed.

```
val dataFrame = glueContext.getCatalogSource(
     database = "<your_database_name>",
     tableName = "<your_table_name>"
   ).getDataFrame()
```

4. Write to a Hudi table registered in Lake Formation. The code is same as writing to a non-registered Hudi table. Note that the AWS Glue job IAM role needs to have the SUPER permission for the write to succeed.

```
glueContext.getCatalogSink("<your_database_name>", "<your_table_name>",
      additionalOptions = JsonOptions(Map(
        "hoodie.table.name" -> "<your_table_name>",
        "hoodie.database.name" -> "<your_database_name>",
        "hoodie.datasource.write.storage.type" -> "COPY_ON_WRITE",
        "hoodie.datasource.write.operation" -> "<write_operation>",
        "hoodie.datasource.write.recordkey.field" -> "<your_recordkey_field>",
        "hoodie.datasource.write.precombine.field" -> "<your_precombine_field>",
        "hoodie.datasource.write.partitionpath.field" -> "<your_partitionkey_field>",
        "hoodie.datasource.write.hive_style_partitioning" -> "true",
        "hoodie.datasource.hive_sync.enable" -> "true",
        "hoodie.datasource.hive_sync.database" -> "<your_database_name>",
        "hoodie.datasource.hive_sync.table" -> "<your_table_name>",
        "hoodie.datasource.hive_sync.partition_fields" ->
 "<your_partitionkey_field>",
        "hoodie.datasource.hive_sync.partition_extractor_class" ->
 "org.apache.hudi.hive.MultiPartKeysValueExtractor",
        "hoodie.datasource.hive_sync.use_jdbc" -> "false",
        "hoodie.datasource.hive_sync.mode" -> "hms"
      )))
      .writeDataFrame(dataFrame, glueContext)
```

Using the Delta Lake framework in AWS Glue

AWS Glue 3.0 and later supports the Linux Foundation Delta Lake framework. Delta Lake is an open-source data lake storage framework that helps you perform ACID transactions, scale metadata handling, and unify streaming and batch data processing. This topic covers available features for using your data in AWS Glue when you transport or store your data in a Delta Lake table. To learn more about Delta Lake, see the official Delta Lake documentation.

You can use AWS Glue to perform read and write operations on Delta Lake tables in Amazon S3, or work with Delta Lake tables using the AWS Glue Data Catalog. Additional operations such as insert, update, and Table batch reads and writes are also supported. When you use Delta Lake tables, you also have the option to use methods from the Delta Lake Python library such as `DeltaTable.forPath`. For more information about the Delta Lake Python library, see Delta Lake's Python documentation.

The following table lists the version of Delta Lake included in each AWS Glue version.

AWS Glue version	Supported Delta Lake version
5.0	3.3.0
4.0	2.1.0
3.0	1.0.0

To learn more about the data lake frameworks that AWS Glue supports, see Using data lake frameworks with AWS Glue ETL jobs.

Enabling Delta Lake for AWS Glue

To enable Delta Lake for AWS Glue, complete the following tasks:

- Specify `delta` as a value for the `--datalake-formats` job parameter. For more information, see Using job parameters in AWS Glue jobs.
- Create a key named `--conf` for your AWS Glue job, and set it to the following value. Alternatively, you can set the following configuration using `SparkConf` in your script. These settings help Apache Spark correctly handle Delta Lake tables.

```
spark.sql.extensions=io.delta.sql.DeltaSparkSessionExtension --conf
 spark.sql.catalog.spark_catalog=org.apache.spark.sql.delta.catalog.DeltaCatalog --
conf
 spark.delta.logStore.class=org.apache.spark.sql.delta.storage.S3SingleDriverLogStore
```

- Lake Formation permission support for Delta tables is enabled by default for AWS Glue 4.0. No additional configuration is needed for reading/writing to Lake Formation-registered Delta tables. To read a registered Delta table, the AWS Glue job IAM role must have the SELECT permission. To write to a registered Delta table, the AWS Glue job IAM role must have the SUPER permission. To learn more about managing Lake Formation permissions, see Granting and revoking permissions on Data Catalog resources.

Using a different Delta Lake version

To use a version of Delta lake that AWS Glue doesn't support, specify your own Delta Lake JAR files using the `--extra-jars` job parameter. Do not include `delta` as a value for the `--datalake-formats` job parameter. If you use AWS Glue 5.0, you must set `--user-jars-first true` job parameter. To use the Delta Lake Python library in this case, you must specify the library JAR files using the `--extra-py-files` job parameter. The Python library comes packaged in the Delta Lake JAR files.

Example: Write a Delta Lake table to Amazon S3 and register it to the AWS Glue Data Catalog

The following AWS Glue ETL script demonstrates how to write a Delta Lake table to Amazon S3 and register the table to the AWS Glue Data Catalog.

Python

```
# Example: Create a Delta Lake table from a DataFrame
# and register the table to Glue Data Catalog

additional_options = {
    "path": "s3://<s3Path>"
}
dataFrame.write \
    .format("delta") \
    .options(**additional_options) \
    .mode("append") \
    .partitionBy("<your_partitionkey_field>") \
    .saveAsTable("<your_database_name>.<your_table_name>")
```

Scala

```
// Example: Example: Create a Delta Lake table from a DataFrame
// and register the table to Glue Data Catalog

val additional_options = Map(
  "path" -> "s3://<s3Path>"
)
dataFrame.write.format("delta")
  .options(additional_options)
  .mode("append")
  .partitionBy("<your_partitionkey_field>")
  .saveAsTable("<your_database_name>.<your_table_name>")
```

Example: Read a Delta Lake table from Amazon S3 using the AWS Glue Data Catalog

The following AWS Glue ETL script reads the Delta Lake table that you created in Example: Write a Delta Lake table to Amazon S3 and register it to the AWS Glue Data Catalog.

Python

For this example, use the create_data_frame.from_catalog method.

```
# Example: Read a Delta Lake table from Glue Data Catalog

from awsglue.context import GlueContext
from pyspark.context import SparkContext

sc = SparkContext()
glueContext = GlueContext(sc)

df = glueContext.create_data_frame.from_catalog(
    database="<your_database_name>",
    table_name="<your_table_name>",
    additional_options=additional_options
)
```

Scala

For this example, use the getCatalogSource method.

```
// Example: Read a Delta Lake table from Glue Data Catalog

import com.amazonaws.services.glue.GlueContext
import org.apacke.spark.SparkContext

object GlueApp {
  def main(sysArgs: Array[String]): Unit = {
    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)
    val df = glueContext.getCatalogSource("<your_database_name>",
 "<your_table_name>",
      additionalOptions = additionalOptions)
      .getDataFrame()
  }
}
```

Example: Insert a `DataFrame` into a Delta Lake table in Amazon S3 using the AWS Glue Data Catalog

This example inserts data into the Delta Lake table that you created in Example: Write a Delta Lake table to Amazon S3 and register it to the AWS Glue Data Catalog.

> **Note**
>
> This example requires you to set the `--enable-glue-datacatalog` job parameter in order to use the AWS Glue Data Catalog as an Apache Spark Hive metastore. To learn more, see Using job parameters in AWS Glue jobs.

Python

For this example, use the write_data_frame.from_catalog method.

```
# Example: Insert into a Delta Lake table in S3 using Glue Data Catalog

from awsglue.context import GlueContext
from pyspark.context import SparkContext

sc = SparkContext()
glueContext = GlueContext(sc)

glueContext.write_data_frame.from_catalog(
    frame=dataFrame,
    database="<your_database_name>",
    table_name="<your_table_name>",
    additional_options=additional_options
)
```

Scala

For this example, use the getCatalogSink method.

```
// Example: Insert into a Delta Lake table in S3 using Glue Data Catalog

import com.amazonaws.services.glue.GlueContext
import org.apacke.spark.SparkContext
```

```
object GlueApp {
  def main(sysArgs: Array[String]): Unit = {
    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)
    glueContext.getCatalogSink("<your_database_name>", "<your_table_name>",
      additionalOptions = additionalOptions)
      .writeDataFrame(dataFrame, glueContext)
  }
}
```

Example: Read a Delta Lake table from Amazon S3 using the Spark API

This example reads a Delta Lake table from Amazon S3 using the Spark API.

Python

```
# Example: Read a Delta Lake table from S3 using a Spark DataFrame

dataFrame = spark.read.format("delta").load("s3://<s3path/>")
```

Scala

```
// Example: Read a Delta Lake table from S3 using a Spark DataFrame

val dataFrame = spark.read.format("delta").load("s3://<s3path/>")
```

Example: Write a Delta Lake table to Amazon S3 using Spark

This example writes a Delta Lake table to Amazon S3 using Spark.

Python

```
# Example: Write a Delta Lake table to S3 using a Spark DataFrame

dataFrame.write.format("delta") \
    .options(**additional_options) \
    .mode("overwrite") \
    .partitionBy("<your_partitionkey_field>")
    .save("s3://<s3Path>")
```

Scala

```
// Example: Write a Delta Lake table to S3 using a Spark DataFrame

dataFrame.write.format("delta")
  .options(additionalOptions)
  .mode("overwrite")
  .partitionBy("<your_partitionkey_field>")
  .save("s3://<s3path/>")
```

Example: Read and write Delta Lake table with Lake Formation permission control

This example reads and writes a Delta Lake table with Lake Formation permission control.

1. Create a Delta table and register it in Lake Formation
   a. To enable Lake Formation permission control, you'll first need to register the table Amazon S3 path on Lake Formation. For more information, see Registering an Amazon S3 location. You can register it either from the Lake Formation console or by using the AWS CLI:

   ```
   aws lakeformation register-resource --resource-arn arn:aws:s3:::<s3-bucket>/<s3-
   folder> --use-service-linked-role --region <REGION>
   ```

   Once you register an Amazon S3 location, any AWS Glue table pointing to the location (or any of its child locations) will return the value for the `IsRegisteredWithLakeFormation` parameter as true in the `GetTable` call.

   b. Create a Delta table that points to the registered Amazon S3 path through Spark:

   > **Note**
   >
   > The following are Python examples.

   ```
   dataFrame.write \
    .format("delta") \
    .mode("overwrite") \
    .partitionBy("<your_partitionkey_field>") \
    .save("s3://<the_s3_path>")
   ```

After the data has been written to Amazon S3, use the AWS Glue crawler to create a new Delta catalog table. For more information, see Introducing native Delta Lake table support with AWS Glue crawlers.

You can also create the table manually through the AWS Glue `CreateTable` API.

2. Grant Lake Formation permission to the AWS Glue job IAM role. You can either grant permissions from the Lake Formation console, or using the AWS CLI. For more information, see Granting table permissions using the Lake Formation console and the named resource method
3. Read the Delta table registered in Lake Formation. The code is same as reading a non-registered Delta table. Note that the AWS Glue job IAM role needs to have the SELECT permission for the read to succeed.

```
# Example: Read a Delta Lake table from Glue Data Catalog

df = glueContext.create_data_frame.from_catalog(
    database="<your_database_name>",
    table_name="<your_table_name>",
    additional_options=additional_options
)
```

4. Write to a Delta table registered in Lake Formation. The code is same as writing to a non-registered Delta table. Note that the AWS Glue job IAM role needs to have the SUPER permission for the write to succeed.

By default AWS Glue uses `Append` as saveMode. You can change it by setting the saveMode option in `additional_options`. For information about saveMode support in Delta tables, see Write to a table.

```
glueContext.write_data_frame.from_catalog(
    frame=dataFrame,
    database="<your_database_name>",
    table_name="<your_table_name>",
    additional_options=additional_options
)
```

Using the Iceberg framework in AWS Glue

AWS Glue 3.0 and later supports the Apache Iceberg framework for data lakes. Iceberg provides a high-performance table format that works just like a SQL table. This topic covers available features for using your data in AWS Glue when you transport or store your data in an Iceberg table. To learn more about Iceberg, see the official Apache Iceberg documentation.

You can use AWS Glue to perform read and write operations on Iceberg tables in Amazon S3, or work with Iceberg tables using the AWS Glue Data Catalog. Additional operations including insert and all Spark Queries Spark Writes are also supported. Update is not supported for Iceberg tables.

> **Note**
>
> `ALTER TABLE … RENAME TO` is not available for Apache Iceberg 0.13.1 for AWS Glue 3.0.

The following table lists the version of Iceberg included in each AWS Glue version.

AWS Glue version	Supported Iceberg version
5.0	1.7.1
4.0	1.0.0
3.0	0.13.1

To learn more about the data lake frameworks that AWS Glue supports, see Using data lake frameworks with AWS Glue ETL jobs.

Enabling the Iceberg framework

To enable Iceberg for AWS Glue, complete the following tasks:

- Specify `iceberg` as a value for the `--datalake-formats` job parameter. For more information, see Using job parameters in AWS Glue jobs.
- Create a key named `--conf` for your AWS Glue job, and set it to the following value. Alternatively, you can set the following configuration using `SparkConf` in your script. These settings help Apache Spark correctly handle Iceberg tables.

```
spark.sql.extensions=org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions
--conf spark.sql.catalog.glue_catalog=org.apache.iceberg.spark.SparkCatalog
--conf spark.sql.catalog.glue_catalog.warehouse=s3://<your-warehouse-dir>/
--conf spark.sql.catalog.glue_catalog.catalog-
impl=org.apache.iceberg.aws.glue.GlueCatalog
--conf spark.sql.catalog.glue_catalog.io-impl=org.apache.iceberg.aws.s3.S3FileIO
```

If you are reading or writing to Iceberg tables that are registered with Lake Formation, follow the guidance in the section called “Lake Formation for FGAC” in AWS Glue 5.0 and later. In AWS Glue 4.0, add the following configuration to enable Lake Formation support.

```
--conf spark.sql.catalog.glue_catalog.glue.lakeformation-enabled=true
--conf spark.sql.catalog.glue_catalog.glue.id=<table-catalog-id>
```

If you use AWS Glue 3.0 with Iceberg 0.13.1, you must set the following additional configurations to use Amazon DynamoDB lock manager to ensure atomic transaction. AWS Glue 4.0 or later uses optimistic locking by default. For more information, see Iceberg AWS Integrations in the official Apache Iceberg documentation.

```
--conf spark.sql.catalog.glue_catalog.lock-
impl=org.apache.iceberg.aws.glue.DynamoLockManager
--conf spark.sql.catalog.glue_catalog.lock.table=<your-dynamodb-table-name>
```

Using a different Iceberg version

To use a version of Iceberg that AWS Glue doesn't support, specify your own Iceberg JAR files using the `--extra-jars` job parameter. Do not include `iceberg` as a value for the `--datalake-formats` parameter. If you use AWS Glue 5.0, you must set `--user-jars-first true` job parameter.

Enabling encryption for Iceberg tables

> **Note**
>
> Iceberg tables have their own mechanisms to enable server-side encryption. You should enable this configuration in addition to AWS Glue's security configuration.

To enable server-side encryption on Iceberg tables, review the guidance from the Iceberg documentation.

Add Spark configuration for Iceberg cross region

To add additional spark configuration for Iceberg cross-region table access with the AWS Glue Data Catalog and AWS Lake Formation, follow the steps below:

1. Create a Multi-region access point.
2. Set the following Spark properties:

```
-----
    --conf spark.sql.catalog.my_catalog.s3.use-arn-region-enabled=true \
    --conf spark.sql.catalog.{CATALOG}.s3.access-points.bucket1",
 "arn:aws:s3::<account-id>:accesspoint/<mrap-id>.mrap \
    --conf spark.sql.catalog.{CATALOG}.s3.access-points.bucket2",
 "arn:aws:s3::<account-id>:accesspoint/<mrap-id>.mrap
-----
```

Example: Write an Iceberg table to Amazon S3 and register it to the AWS Glue Data Catalog

This example script demonstrates how to write an Iceberg table to Amazon S3. The example uses Iceberg AWS Integrations to register the table to the AWS Glue Data Catalog.

Python

```
# Example: Create an Iceberg table from a DataFrame
# and register the table to Glue Data Catalog

dataFrame.createOrReplaceTempView("tmp_<your_table_name>")

query = f"""
CREATE TABLE glue_catalog.<your_database_name>.<your_table_name>
USING iceberg
TBLPROPERTIES ("format-version"="2")
AS SELECT * FROM tmp_<your_table_name>
"""
spark.sql(query)
```

Scala

```
// Example: Example: Create an Iceberg table from a DataFrame
// and register the table to Glue Data Catalog

dataFrame.createOrReplaceTempView("tmp_<your_table_name>")

val query = """CREATE TABLE glue_catalog.<your_database_name>.<your_table_name>
USING iceberg
TBLPROPERTIES ("format-version"="2")
AS SELECT * FROM tmp_<your_table_name>
"""
spark.sql(query)
```

Alternatively, you can write an Iceberg table to Amazon S3 and the Data Catalog using Spark methods.

Prerequisites: You will need to provision a catalog for the Iceberg library to use. When using the AWS Glue Data Catalog, AWS Glue makes this straightforward. The AWS Glue Data Catalog is pre-configured for use by the Spark libraries as `glue_catalog`. Data Catalog tables are identified by a *`databaseName`* and a *`tableName`*. For more information about the AWS Glue Data Catalog, see *Data discovery and cataloging*.

If you are not using the AWS Glue Data Catalog, you will need to provision a catalog through the Spark APIs. For more information, see Spark Configuration in the Iceberg documentation.

This example writes an Iceberg table to Amazon S3 and the Data Catalog using Spark.

Python

```
# Example: Write an Iceberg table to S3 on the Glue Data Catalog

# Create (equivalent to CREATE TABLE AS SELECT)
dataFrame.writeTo("glue_catalog.databaseName.tableName") \
    .tableProperty("format-version", "2") \
    .create()

# Append (equivalent to INSERT INTO)
dataFrame.writeTo("glue_catalog.databaseName.tableName") \
    .tableProperty("format-version", "2") \
    .append()
```

Scala

```
// Example: Write an Iceberg table to S3 on the Glue Data Catalog

// Create (equivalent to CREATE TABLE AS SELECT)
dataFrame.writeTo("glue_catalog.databaseName.tableName")
    .tableProperty("format-version", "2")
    .create()

// Append (equivalent to INSERT INTO)
dataFrame.writeTo("glue_catalog.databaseName.tableName")
    .tableProperty("format-version", "2")
    .append()
```

Example: Read an Iceberg table from Amazon S3 using the AWS Glue Data Catalog

This example reads the Iceberg table that you created in Example: Write an Iceberg table to Amazon S3 and register it to the AWS Glue Data Catalog.

Python

For this example, use the `GlueContext.create_data_frame.from_catalog()` method.

```
# Example: Read an Iceberg table from Glue Data Catalog

from awsglue.context import GlueContext
from pyspark.context import SparkContext

sc = SparkContext()
glueContext = GlueContext(sc)

df = glueContext.create_data_frame.from_catalog(
    database="<your_database_name>",
    table_name="<your_table_name>",
    additional_options=additional_options
)
```

Scala

For this example, use the getCatalogSource method.

```
// Example: Read an Iceberg table from Glue Data Catalog
```

```
import com.amazonaws.services.glue.GlueContext
import org.apacke.spark.SparkContext

object GlueApp {
  def main(sysArgs: Array[String]): Unit = {
    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)
    val df = glueContext.getCatalogSource("<your_database_name>",
 "<your_table_name>",
      additionalOptions = additionalOptions)
      .getDataFrame()
  }
}
```

Example: Insert a `DataFrame` into an Iceberg table in Amazon S3 using the AWS Glue Data Catalog

This example inserts data into the Iceberg table that you created in Example: Write an Iceberg table to Amazon S3 and register it to the AWS Glue Data Catalog.

> **Note**
>
> This example requires you to set the `--enable-glue-datacatalog` job parameter in order to use the AWS Glue Data Catalog as an Apache Spark Hive metastore. To learn more, see Using job parameters in AWS Glue jobs.

Python

For this example, use the `GlueContext.write_data_frame.from_catalog()` method.

```
# Example: Insert into an Iceberg table from Glue Data Catalog

from awsglue.context import GlueContext
from pyspark.context import SparkContext

sc = SparkContext()
glueContext = GlueContext(sc)

glueContext.write_data_frame.from_catalog(
    frame=dataFrame,
    database="<your_database_name>",
    table_name="<your_table_name>",
    additional_options=additional_options
)
```

Scala

For this example, use the getCatalogSink method.

```
// Example: Insert into an Iceberg table from Glue Data Catalog

import com.amazonaws.services.glue.GlueContext
import org.apacke.spark.SparkContext

object GlueApp {
  def main(sysArgs: Array[String]): Unit = {
    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)
    glueContext.getCatalogSink("<your_database_name>", "<your_table_name>",
      additionalOptions = additionalOptions)
      .writeDataFrame(dataFrame, glueContext)
  }
}
```

Example: Read an Iceberg table from Amazon S3 using Spark

Prerequisites: You will need to provision a catalog for the Iceberg library to use. When using the AWS Glue Data Catalog, AWS Glue makes this straightforward. The AWS Glue Data Catalog is pre-configured for use by the Spark libraries as `glue_catalog`. Data Catalog tables are identified by a *`databaseName`* and a *`tableName`*. For more information about the AWS Glue Data Catalog, see *Data discovery and cataloging*.

If you are not using the AWS Glue Data Catalog, you will need to provision a catalog through the Spark APIs. For more information, see Spark Configuration in the Iceberg documentation.

This example reads an Iceberg table in Amazon S3 from the Data Catalog using Spark.

Python

```
# Example: Read an Iceberg table on S3 as a DataFrame from the Glue Data Catalog

```

```
dataFrame = spark.read.format("iceberg").load("glue_catalog.databaseName.tableName")
```

Scala

```
// Example: Read an Iceberg table on S3 as a DataFrame from the Glue Data Catalog

val dataFrame =
 spark.read.format("iceberg").load("glue_catalog.databaseName.tableName")
```

Example: Read and write Iceberg table with Lake Formation permission control

This example reads and writes an Iceberg table with Lake Formation permission control.

> **Note**
>
> This example works only in AWS Glue 4.0. In AWS Glue 5.0 and later, follow the guidance in the section called “Lake Formation for FGAC”.

1. Create an Iceberg table and register it in Lake Formation:
   a. To enable Lake Formation permission control, you’ll first need to register the table Amazon S3 path on Lake Formation. For more information, see Registering an Amazon S3 location. You can register it either from the Lake Formation console or by using the AWS CLI:

      ```
      aws lakeformation register-resource --resource-arn arn:aws:s3:::<s3-bucket>/<s3-
      folder> --use-service-linked-role --region <REGION>
      ```

      Once you register an Amazon S3 location, any AWS Glue table pointing to the location (or any of its child locations) will return the value for the `IsRegisteredWithLakeFormation` parameter as true in the `GetTable` call.

   b. Create an Iceberg table that points to the registered path through Spark SQL:

      > **Note**
      >
      > The following are Python examples.

      ```
      dataFrame.createOrReplaceTempView("tmp_<your_table_name>")
      ```

```
query = f"""
CREATE TABLE glue_catalog.<your_database_name>.<your_table_name>
USING iceberg
AS SELECT * FROM tmp_<your_table_name>
"""
spark.sql(query)
```

You can also create the table manually through AWS Glue `CreateTable` API. For more information, see [Creating Apache Iceberg tables](https://docs.aws.amazon.com/).

> **Note**
>
> The `UpdateTable` API does not currently support Iceberg table format as an input to the operation.

2. Grant Lake Formation permission to the job IAM role. You can either grant permissions from the Lake Formation console, or using the AWS CLI. For more information, see: https://docs.aws.amazon.com/lake-formation/latest/dg/granting-table-permissions.html
3. Read an Iceberg table registered with Lake Formation. The code is same as reading a non-registered Iceberg table. Note that your AWS Glue job IAM role needs to have the SELECT permission for the read to succeed.

```
# Example: Read an Iceberg table from the AWS Glue Data Catalog
from awsglue.context import GlueContextfrom pyspark.context import SparkContext

sc = SparkContext()
glueContext = GlueContext(sc)

df = glueContext.create_data_frame.from_catalog(
    database="<your_database_name>",
    table_name="<your_table_name>",
    additional_options=additional_options
)
```

4. Write to an Iceberg table registered with Lake Formation. The code is same as writing to a non-registered Iceberg table. Note that your AWS Glue job IAM role needs to have the SUPER permission for the write to succeed.

```
glueContext.write_data_frame.from_catalog(
    frame=dataFrame,
    database="<your_database_name>",
    table_name="<your_table_name>",
    additional_options=additional_options
)
```

Shared configuration reference

You can use the following `format_options` values with any format type.

- `attachFilename` — A string in the appropriate format to be used as a column name. If you provide this option, the name of the source file for the record will be appended to the record. The parameter value will be used as the column name.
- `attachTimestamp` — A string in the appropriate format to be used as a column name. If you provide this option, the modification time of the source file for the record will be appended to the record. The parameter value will be used as the column name.

AWS Glue Data Catalog support for Spark SQL jobs

The AWS Glue Data Catalog is an Apache Hive metastore-compatible catalog. You can configure your AWS Glue jobs and development endpoints to use the Data Catalog as an external Apache Hive metastore. You can then directly run Apache Spark SQL queries against the tables stored in the Data Catalog. AWS Glue dynamic frames integrate with the Data Catalog by default. However, with this feature, Spark SQL jobs can start using the Data Catalog as an external Hive metastore.

This feature requires network access to the AWS Glue API endpoint. For AWS Glue jobs with connections located in private subnets, you must configure either a VPC endpoint or NAT gateway to provide the network access. For information about configuring a VPC endpoint, see Setting up network access to data stores. To create a NAT gateway, see NAT Gateways in the *Amazon VPC User Guide*.

You can configure AWS Glue jobs and development endpoints by adding the `"--enable-glue-datacatalog": ""` argument to job arguments and development endpoint arguments respectively. Passing this argument sets certain configurations in Spark that enable it to access the Data Catalog as an external Hive metastore. It also enables Hive support in the `SparkSession` object created in the AWS Glue job or development endpoint.

To enable the Data Catalog access, check the **Use AWS Glue Data Catalog as the Hive metastore** check box in the **Catalog options** group on the **Add job** or **Add endpoint** page on the console. Note that the IAM role used for the job or development endpoint should have `glue:CreateDatabase` permissions. A database called "`default`" is created in the Data Catalog if it does not exist.

Lets look at an example of how you can use this feature in your Spark SQL jobs. The following example assumes that you have crawled the US legislators dataset available at `s3://awsglue-datasets/examples/us-legislators`.

To serialize/deserialize data from the tables defined in the AWS Glue Data Catalog, Spark SQL needs the Hive SerDe class for the format defined in the AWS Glue Data Catalog in the classpath of the spark job.

SerDes for certain common formats are distributed by AWS Glue. The following are the Amazon S3 links for these:

- JSON
- XML
- Grok

Add the JSON SerDe as an extra JAR to the development endpoint. For jobs, you can add the SerDe using the `--extra-jars` argument in the arguments field. For more information, see Using job parameters in AWS Glue jobs.

Here is an example input JSON to create a development endpoint with the Data Catalog enabled for Spark SQL.

```
{
    "EndpointName": "Name",
    "RoleArn": "role_ARN",
    "PublicKey": "public_key_contents",
    "NumberOfNodes": 2,
    "Arguments": {
      "--enable-glue-datacatalog": ""
    },
    "ExtraJarsS3Path": "s3://crawler-public/json/serde/json-serde.jar"
}
```

Now query the tables created from the US legislators dataset using Spark SQL.

```
>>> spark.sql("use legislators")
DataFrame[]
>>> spark.sql("show tables").show()
+-----------+------------------+-----------+
|   database|         tableName|isTemporary|
+-----------+------------------+-----------+
|legislators|        areas_json|      false|
|legislators|    countries_json|      false|
|legislators|       events_json|      false|
|legislators|  memberships_json|      false|
|legislators|organizations_json|      false|
|legislators|      persons_json|      false|
+-----------+------------------+-----------+
>>> spark.sql("describe memberships_json").show()
+--------------------+---------+-----------------+
|            col_name|data_type|          comment|
+--------------------+---------+-----------------+
|             area_id|   string|from deserializer|
|     on_behalf_of_id|   string|from deserializer|
|     organization_id|   string|from deserializer|
|                role|   string|from deserializer|
|           person_id|   string|from deserializer|
|legislative_perio...|   string|from deserializer|
|          start_date|   string|from deserializer|
|            end_date|   string|from deserializer|
+--------------------+---------+-----------------+
```

If the SerDe class for the format is not available in the job's classpath, you will see an error similar to the following.

```
>>> spark.sql("describe memberships_json").show()

Caused by: MetaException(message:java.lang.ClassNotFoundException Class
 org.openx.data.jsonserde.JsonSerDe not found)
    at
 org.apache.hadoop.hive.metastore.MetaStoreUtils.getDeserializer(MetaStoreUtils.java:399)
    at
 org.apache.hadoop.hive.ql.metadata.Table.getDeserializerFromMetaStore(Table.java:276)
    ... 64 more
```

To view only the distinct `organization_id`s from the `memberships` table, run the following SQL query.

```
>>> spark.sql("select distinct organization_id from memberships_json").show()
+--------------------+
|     organization_id|
+--------------------+
|d56acebe-8fdc-47b...|
|8fa6c3d2-71dc-478...|
+--------------------+
```

If you need to do the same with dynamic frames, run the following.

```
>>> memberships = glueContext.create_dynamic_frame.from_catalog(database="legislators",
 table_name="memberships_json")
>>> memberships.toDF().createOrReplaceTempView("memberships")
>>> spark.sql("select distinct organization_id from memberships").show()
+--------------------+
|     organization_id|
+--------------------+
|d56acebe-8fdc-47b...|
|8fa6c3d2-71dc-478...|
+--------------------+
```

While DynamicFrames are optimized for ETL operations, enabling Spark SQL to access the Data Catalog directly provides a concise way to run complex SQL statements or port existing applications.

Using job bookmarks

AWS Glue for Spark uses job bookmarks to track data that has already been processed. For a summary of the job bookmarks feature and what it supports, see the section called “Tracking processed data using job bookmarks”. When programming a AWS Glue job with bookmarks, you have access to flexibility unavailable in visual jobs.

- When reading from JDBC, you can specify the column(s) to use as bookmark keys in your AWS Glue script.
- You can chose which `transformation_ctx` to apply to each method call.

Always call `job.init` in the beginning of the script and the `job.commit` in the end of the script with appropriately configured parameters. These two functions initialize the bookmark service and update the state change to the service. Bookmarks won't work without calling them.

Specify bookmark keys

For JDBC workflows, the bookmark keeps track of which rows your job has read by comparing the values of key fields to a bookmarked value. This is not necessary or applicable for Amazon S3 workflows. When writing a AWS Glue script without the visual editor, you can specify which column to track with bookmarks. You can also specify multiple columns. Gaps in the sequence of values are permitted when specifying user-defined bookmark keys.

> **Warning**
>
> If user-defined bookmarks keys are used, they must each be strictly monotonically increasing or decreasing. When selecting additional fields for a compound key, fields for concepts like "minor versions" or "revision numbers" do not meet this criteria, since their values are reused throughout your dataset.

You can specify `jobBookmarkKeys` and `jobBookmarkKeysSortOrder` in the following ways:

- `create_dynamic_frame.from_catalog` — Use `additional_options`.
- `create_dynamic_frame.from_options` — Use `connection_options`.

Transformation context

Many of the AWS Glue PySpark dynamic frame methods include an optional parameter named `transformation_ctx`, which is a unique identifier for the ETL operator instance. The `transformation_ctx` parameter is used to identify state information within a job bookmark for the given operator. Specifically, AWS Glue uses `transformation_ctx` to index the key to the bookmark state.

> **Warning**
>
> The `transformation_ctx` serves as the key to search the bookmark state for a specific source in your script. For the bookmark to work properly, you should always keep the source and the associated `transformation_ctx` consistent. Changing the source

property or renaming the `transformation_ctx` may make the previous bookmark invalid and the time stamp based filtering may not yield the correct result.

For job bookmarks to work properly, enable the job bookmark parameter and set the `transformation_ctx` parameter. If you don't pass in the `transformation_ctx` parameter, then job bookmarks are not enabled for a dynamic frame or a table used in the method. For example, if you have an ETL job that reads and joins two Amazon S3 sources, you might choose to pass the `transformation_ctx` parameter only to those methods that you want to enable bookmarks. If you reset the job bookmark for a job, it resets all transformations that are associated with the job regardless of the `transformation_ctx` used.

For more information about the `DynamicFrameReader` class, see DynamicFrameReader class. For more information about PySpark extensions, see AWS Glue PySpark extensions reference.

Examples

Example

The following is an example of a generated script for an Amazon S3 data source. The portions of the script that are required for using job bookmarks are shown in italics. For more information about these elements see the GlueContext class API, and the DynamicFrameWriter class API.

```
# Sample Script
import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job

args = getResolvedOptions(sys.argv, ['JOB_NAME'])
sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
job.init(args['JOB_NAME'], args)

datasource0 = glueContext.create_dynamic_frame.from_catalog(
    database = "database",
    table_name = "relatedqueries_csv",
    transformation_ctx = "datasource0"
)

applymapping1 = ApplyMapping.apply(
    frame = datasource0,
    mappings = [("col0", "string", "name", "string"), ("col1", "string", "number",
 "string")],
    transformation_ctx = "applymapping1"
)

datasink2 = glueContext.write_dynamic_frame.from_options(
    frame = applymapping1,
    connection_type = "s3",
    connection_options = {"path": "s3://input_path"},
    format = "json",
    transformation_ctx = "datasink2"
)


job.commit()
```

Example

The following is an example of a generated script for a JDBC source. The source table is an employee table with the `empno` column as the primary key. Although by default the job uses a sequential primary key as the bookmark key if no bookmark key is specified, because `empno` is not necessarily sequential—there could be gaps in the values—it does not qualify as a default bookmark key. Therefore, the script explicitly designates `empno` as the bookmark key. That portion of the code is shown in italics.

```
import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job

args = getResolvedOptions(sys.argv, ['JOB_NAME'])

sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
job.init(args['JOB_NAME'], args)
```

```
datasource0 = glueContext.create_dynamic_frame.from_catalog(
    database = "hr",
    table_name = "emp",
    transformation_ctx = "datasource0",
    additional_options = {"jobBookmarkKeys":["empno"],"jobBookmarkKeysSortOrder":"asc"}
)

applymapping1 = ApplyMapping.apply(
    frame = datasource0,
    mappings = [("ename", "string", "ename", "string"), ("hrly_rate", "decimal(38,0)",
 "hrly_rate", "decimal(38,0)"), ("comm", "decimal(7,2)", "comm", "decimal(7,2)"),
 ("hiredate", "timestamp", "hiredate", "timestamp"), ("empno", "decimal(5,0)", "empno",
 "decimal(5,0)"), ("mgr", "decimal(5,0)", "mgr", "decimal(5,0)"), ("photo", "string",
 "photo", "string"), ("job", "string", "job", "string"), ("deptno", "decimal(3,0)",
 "deptno", "decimal(3,0)"), ("ssn", "decimal(9,0)", "ssn", "decimal(9,0)"), ("sal",
 "decimal(7,2)", "sal", "decimal(7,2)")],
    transformation_ctx = "applymapping1"
)

datasink2 = glueContext.write_dynamic_frame.from_options(
    frame = applymapping1,
    connection_type = "s3",
    connection_options = {"path": "s3://hr/employees"},
    format = "csv",
    transformation_ctx = "datasink2"
)

job.commit()
```

Using Sensitive Data Detection outside AWS Glue Studio

AWS Glue Studio allows you to detect sensitive data, however, you can also use the Sensitive Data Detection functionality outside of AWS Glue Studio.

For a full list of managed sensitive data types, see Managed data types.

Detecting Sensitive Data Detection using AWS Managed PII types

AWS Glue provides two APIs in a AWS Glue ETL job. These are `detect()` and `classifyColumns()`:

```
 detect(frame: DynamicFrame,
     entityTypesToDetect: Seq[String],
     outputColumnName: String = "DetectedEntities",
     detectionSensitivity: String = "LOW"): DynamicFrame

detect(frame: DynamicFrame,
     detectionParameters: JsonOptions,
     outputColumnName: String = "DetectedEntities",
     detectionSensitivity: String = "LOW"): DynamicFrame

 classifyColumns(frame: DynamicFrame,
     entityTypesToDetect: Seq[String],
     sampleFraction: Double = 0.1,
     thresholdFraction: Double = 0.1,
     detectionSensitivity: String = "LOW")
```

You can use the `detect()` API to identify AWS Managed PII types and custom entity types. A new column is automatically created with the detection result. The `classifyColumns()` API returns a map where keys are column names and values are list of detected entity types. `SampleFraction` indicates the fraction of the data to sample when scanning for PII entities whereas `ThresholdFraction` indicates the fraction of the data that must be met in order for a column to be identified as PII data.

Row-level detection

In the example, the job is performing the following actions using the `detect()` and `classifyColumns()` APIs:

- reading data from an Amazon S3 bucket and turns it into a dynamicFrame
- detecting instances of "Email" and "Credit Card" in the dynamicFrame
- returning a dynamicFrame with original values plus one column which encompasses detection result for each row
- writing the returned dynamicFrame in another Amazon S3 path

```
import com.amazonaws.services.glue.GlueContext
import com.amazonaws.services.glue.MappingSpec
import com.amazonaws.services.glue.errors.CallSite
```

```
 import com.amazonaws.services.glue.util.GlueArgParser
 import com.amazonaws.services.glue.util.Job
 import com.amazonaws.services.glue.util.JsonOptions
 import org.apache.spark.SparkContext
 import scala.collection.JavaConverters._
 import com.amazonaws.services.glue.ml.EntityDetector

 object GlueApp {
   def main(sysArgs: Array[String]) {
     val spark: SparkContext = new SparkContext()
     val glueContext: GlueContext = new GlueContext(spark)
     val args = GlueArgParser.getResolvedOptions(sysArgs, Seq("JOB_NAME").toArray)
     Job.init(args("JOB_NAME"), glueContext, args.asJava)
     val frame=
glueContext.getSourceWithFormat(formatOptions=JsonOptions("""{"quoteChar": "\"",
"withHeader": true, "separator": ","}"""), connectionType="s3", format="csv",
options=JsonOptions("""{"paths": ["s3://pathToSource"], "recurse": true}"""),
transformationContext="AmazonS3_node1650160158526").getDynamicFrame()

     val frameWithDetectedPII = EntityDetector.detect(frame, Seq("EMAIL",
"CREDIT_CARD"))

     glueContext.getSinkWithFormat(connectionType="s3",
options=JsonOptions("""{"path": "s3://pathToOutput/", "partitionKeys": []}"""),
transformationContext="someCtx",
format="json").writeDynamicFrame(frameWithDetectedPII)

     Job.commit()
   }
 }
```

Row-level detection with fine-grained actions

In the example, the job is performing the following actions using the `detect()` APIs:

- reading data from an Amazon S3 bucket and turns it into a dynamicFrame
- detecting sensitive data types for “USA_PTIN”, “ BANK_ACCOUNT”, “USA_SSN”, “USA_PASSPORT_NUMBER” , and “PHONE_NUMBER” in the dynamicFrame
- returning a dynamicFrame with modified masked values plus one column which encompasses detection result for each row
- writing the returned dynamicFrame in another Amazon S3 path

In contrast with the above `detect()` API, this uses fine-grained actions for entity types to detect. For more information, see [Detection parameters for using `detect()`](#).

```
import com.amazonaws.services.glue.GlueContext
import com.amazonaws.services.glue.MappingSpec
import com.amazonaws.services.glue.errors.CallSite
import com.amazonaws.services.glue.util.GlueArgParser
import com.amazonaws.services.glue.util.Job
import com.amazonaws.services.glue.util.JsonOptions
import org.apache.spark.SparkContext
import scala.collection.JavaConverters._
import com.amazonaws.services.glue.ml.EntityDetector

object GlueApp {
  def main(sysArgs: Array[String]) {
    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)
    val args = GlueArgParser.getResolvedOptions(sysArgs, Seq("JOB_NAME").toArray)
    Job.init(args("JOB_NAME"), glueContext, args.asJava)
    val frame =
 glueContext.getSourceWithFormat(formatOptions=JsonOptions("""{"quoteChar": "\"",
 "withHeader": true, "separator": ","}"""), connectionType="s3", format="csv",
 options=JsonOptions("""{"paths": ["s3://pathToSource"], "recurse": true}"""),
 transformationContext="AmazonS3_node_source").getDynamicFrame()

    val detectionParameters = JsonOptions(
      """
        {
          "USA_DRIVING_LICENSE": [{
            "action": "PARTIAL_REDACT",
            "sourceColumns": ["Driving License"],
            "actionOptions": {
              "matchPattern": "[0-9]",
              "redactChar": "*"
            }
          }],
          "BANK_ACCOUNT": [{
            "action": "DETECT",
            "sourceColumns": ["*"]
          }],
          "USA_SSN": [{
            "action": "SHA256_HASH",
            "sourceColumns": ["SSN"]
          }],
          "IP_ADDRESS": [{
            "action": "REDACT",
            "sourceColumns": ["IP Address"],
            "actionOptions": {"redactText": "*****"}
          }],
          "PHONE_NUMBER": [{
            "action": "PARTIAL_REDACT",
            "sourceColumns": ["Phone Number"],
            "actionOptions": {
              "numLeftCharsToExclude": 1,
              "numRightCharsToExclude": 0,
              "redactChar": "*"
            }
          }]
        }
      """
    )

    val frameWithDetectedPII = EntityDetector.detect(frame, detectionParameters,
 "DetectedEntities", "HIGH")

    glueContext.getSinkWithFormat(connectionType="s3", options=JsonOptions("""{"path":
 "s3://pathToOutput/", "partitionKeys": []}"""),
 transformationContext="AmazonS3_node_target",
 format="json").writeDynamicFrame(frameWithDetectedPII)

    Job.commit()
  }
}
```

Column-level detection

In the example, the job is performing the following actions using the `classifyColumns()`APIs:

- reading data from an Amazon S3 bucket and turns it into a dynamicFrame
- detecting instances of "Email" and "Credit Card" in the dynamicFrame
- set parameters to sample 100% of the column, mark an entity as detected if it is in 10% of cells, and have “LOW” sensitivity
- returns a map where keys are column names and values are list of detected entity types

- writing the returned dynamicFrame in another Amazon S3 path

```
import com.amazonaws.services.glue.GlueContext
import com.amazonaws.services.glue.MappingSpec
import com.amazonaws.services.glue.errors.CallSite
import com.amazonaws.services.glue.util.GlueArgParser
import com.amazonaws.services.glue.util.Job
import com.amazonaws.services.glue.util.JsonOptions
import org.apache.spark.SparkContext
import scala.collection.JavaConverters._
import com.amazonaws.services.glue.DynamicFrame
import com.amazonaws.services.glue.ml.EntityDetector

object GlueApp {
  def main(sysArgs: Array[String]) {
    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)
    val args = GlueArgParser.getResolvedOptions(sysArgs, Seq("JOB_NAME").toArray)
    Job.init(args("JOB_NAME"), glueContext, args.asJava)
    val frame =
 glueContext.getSourceWithFormat(formatOptions=JsonOptions("""{"quoteChar":
 "\"", "withHeader": true, "separator": ",", "optimizePerformance": false}"""),
 connectionType="s3", format="csv", options=JsonOptions("""{"paths": ["s3://
pathToSource"], "recurse": true}"""), transformationContext="frame").getDynamicFrame()

    import glueContext.sparkSession.implicits._

    val detectedDataFrame = EntityDetector.classifyColumns(
        frame,
        entityTypesToDetect = Seq("CREDIT_CARD", "PHONE_NUMBER"),
        sampleFraction = 1.0,
        thresholdFraction = 0.1,
        detectionSensitivity = "LOW"
    )
    val detectedDF = (detectedDataFrame).toSeq.toDF("columnName", "entityTypes")
    val DetectSensitiveData_node = DynamicFrame(detectedDF, glueContext)

    glueContext.getSinkWithFormat(connectionType="s3", options=JsonOptions("""{"path":
 "s3://pathToOutput", "partitionKeys": []}"""), transformationContext="someCtx",
 format="json").writeDynamicFrame(DetectSensitiveData_node)

    Job.commit()
  }
}
```

Detecting Sensitive Data Detection using AWS CustomEntityType PII types

You can define custom entities through AWS Studio. However, to use this feature out of AWS Studio, you have to first define the custom entity types and then add the defined custom entity types to the list of `entityTypesToDetect`.

If you have specific sensitive data types in your data (such as 'Employee Id'), you can create custom entities by calling the `CreateCustomEntityType()` API. The following example defines the custom entity type 'EMPLOYEE_ID' to the `CreateCustomEntityType()` API with the request parameters:

```
{
    "name": "EMPLOYEE_ID",
    "regexString": "\d{4}-\d{3}",
    "contextWords": ["employee"]
}
```

Then, modify the job to use the new custom sensitive data type by adding the custom entity type (EMPLOYEE_ID) to the `EntityDetector()` API:

```
import com.amazonaws.services.glue.GlueContext
import com.amazonaws.services.glue.MappingSpec
import com.amazonaws.services.glue.errors.CallSite
import com.amazonaws.services.glue.util.GlueArgParser
import com.amazonaws.services.glue.util.Job
import com.amazonaws.services.glue.util.JsonOptions
import org.apache.spark.SparkContext
import scala.collection.JavaConverters._
import com.amazonaws.services.glue.ml.EntityDetector

object GlueApp {
  def main(sysArgs: Array[String]) {
    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)
    val args = GlueArgParser.getResolvedOptions(sysArgs, Seq("JOB_NAME").toArray)
```

```
     Job.init(args("JOB_NAME"), glueContext, args.asJava)
     val frame=
glueContext.getSourceWithFormat(formatOptions=JsonOptions("""{"quoteChar": "\"",
"withHeader": true, "separator": ","}"""), connectionType="s3", format="csv",
options=JsonOptions("""{"paths": ["s3://pathToSource"], "recurse": true}"""),
transformationContext="AmazonS3_node1650160158526").getDynamicFrame()

     val frameWithDetectedPII = EntityDetector.detect(frame, Seq("EMAIL",
"CREDIT_CARD", "EMPLOYEE_ID"))

     glueContext.getSinkWithFormat(connectionType="s3",
options=JsonOptions("""{"path": "s3://pathToOutput/", "partitionKeys": []}"""),
transformationContext="someCtx",
format="json").writeDynamicFrame(frameWithDetectedPII)

     Job.commit()
   }
 }
```

> **Note**
>
> If a custom sensitive data type is defined with the same name as an existing managed entity type, then the custom sensitive data type will take precedent and overwrite the managed entity type's logic.

Detection parameters for using `detect()`

This method is used for detecting entities in a DynamicFrame. It returns a new DataFrame with original values and an additional column outputColumnName that has PII detection metadata. Custom masking can be done after this DynamicFrame is returned within the AWS Glue script, or the detect() with fine-grained actions API can be used instead.

```
detect(frame: DynamicFrame,
       entityTypesToDetect: Seq[String],
       outputColumnName: String = "DetectedEntities",
       detectionSensitivity: String = "LOW"): DynamicFrame
```

Parameters:

- **frame –** (type: `DynamicFrame`) The input DynamicFrame containing the data to be processed.
- **entityTypesToDetect –** (type: `[Seq[String]`) List of entity types to detect. Can be either Managed Entity Types or Custom Entity Types.
- **outputColumnName –** (type: `String`, default: "DetectedEntities") The name of the column where detected entities will be stored. If not provided, the default column name is "DetectedEntities".
- **detectionSensitivity –** (type: `String`, options: "LOW" or "HIGH", default: "LOW") Specifies the sensitivity of the detection process. Valid options are "LOW" or "HIGH". If not provided, the default sensitivity is set to "LOW".

`outputColumnName` settings:

The name of the column where detected entities will be stored. If not provided, the default column name is "DetectedEntities". For each row in the output column, the supplementary column includes a map of the column name to the detected entity metadata with the following key-value pairs:

- **entityType –** The detected entity type.
- **start –** The starting position of the detected entity in the original data.
- **end –** The ending position of the detected entity in the original data.
- **actionUsed –** The action performed on the detected entity (e.g., "DETECT," "REDACT," "PARTIAL_REDACT," "SHA256_HASH").

Example:

```
{
   "DetectedEntities":{
      "SSN Col":[
         {
            "entityType":"USA_SSN",
            "actionUsed":"DETECT",
            "start":4,
            "end":15
         }
      ],
      "Random Data col":[
         {
            "entityType":"BANK_ACCOUNT",
            "actionUsed":"PARTIAL_REDACT",
            "start":4,
            "end":13
         },
         {
            "entityType":"IP_ADDRESS",
            "actionUsed":"REDACT",
            "start":4,
            "end":13
         }
      ]
   }
}
```

Detection Parameters for `detect()` with fine grained actions

This method is used for detecting entities in a DynamicFrame using specified parameters. It returns a new DataFrame with original values replaced with masked sensitive data and an additional column `outputColumnName` that has PII detection metadata.

```
detect(frame: DynamicFrame,
       detectionParameters: JsonOptions,
       outputColumnName: String = "DetectedEntities",
       detectionSensitivity: String = "LOW"): DynamicFrame
```

Parameters:

- **frame** – (type: `DynamicFrame`): The input DynamicFrame containing the data to be processed.
- **detectionParameters** – (type: `JsonOptions`): JSON options specifying parameters for the detection process.
- **outputColumnName** – (type: `String`, default: "DetectedEntities"): The name of the column where detected entities will be stored. If not provided, the default column name is "DetectedEntities".
- **detectionSensitivity** – (type: `String`, options: "LOW" or "HIGH", default: "LOW"): Specifies the sensitivity of the detection process. Valid options are "LOW" or "HIGH". If not provided, the default sensitivity is set to "LOW".

`detectionParameters` settings

If no settings are included, default values will be used.

- **action** – (type: `String`, options: "DETECT", "REDACT", "PARTIAL_REDACT", "SHA256_HASH") Specifies the action to be performed on the entity. Required. Note that actions that perform masking (all but "DETECT") can only perform one action per column. This is a preventative measure for masking coalesced entities.
- **sourceColumns** – (type: `List[String]`, default: [“*”]) List of source column names to perform detection on for the entity. Defaults to [“*”] if not present. Raises `IllegalArgumentException` if an invalid column name is used.
- **sourceColumnsToExclude** – (type: `List[String]`) List of source column names to to perform detection on for the entity. Use either `sourceColumns` or `sourceColumnsToExclude`. Raises `IllegalArgumentException` if an invalid column name is used.
- **actionOptions** – Additional options based on the specified action:
  - For "DETECT" and "SHA256_HASH", no options are allowed.
  - For "REDACT":
    - **redactText** – (type: `String`, default: "*****") Text to replace the detected entity.
  - For "PARTIAL_REDACT":
    - **redactChar** – (type: `String`, default: "*") Character to replace each detected character in the entity.
    - **matchPattern** – (type: `String`) Regex pattern for partial redaction. Cannot be combined with numLeftCharsToExclude or `numRightCharsToExclude`.
    - **numLeftCharsToExclude** – (type: `String, integer`) Number of left characters to exclude. Cannot be combined with matchPattern, but can be used with `numRightCharsToExclude`.
    - **numRightCharsToExclude** – (type: `String, integer`) Number of right characters to exclude. Cannot be combined with matchPattern, but can be used with `numRightCharsToExclude`.

`outputColumnName` settings

[See outputColumnName settings]

Detection Parameters for `classifyColumns()`

This method is used for detecting entities in a DynamicFrame. It returns a map where keys are column names and values are list of detected entity types. Custom masking can be done after this is returned within the AWS Glue script.

```
classifyColumns(frame: DynamicFrame,
                entityTypesToDetect: Seq[String],
                sampleFraction: Double = 0.1,
                thresholdFraction: Double = 0.1,
                detectionSensitivity: String = "LOW")
```

Parameters:

- **frame** – (type: `DynamicFrame`) The input DynamicFrame containing the data to be processed.
- **entityTypesToDetect** – (type: `Seq[String]`) List of entity types to detect. Can be either Managed Entity Types or Custom Entity Types.
- **sampleFraction** – (type: `Double`, default: 10%) The fraction of the data to sample when scanning for PII entities.
- **thresholdFraction** – (type: `Double`, default: 10%): The fraction of the data that must be met in order for a column to be identified as PII data.
- **detectionSensitivity** – (type: `String`, options: "LOW" or "HIGH", default: "LOW") Specifies the sensitivity of the detection process. Valid options are "LOW" or "HIGH". If not provided, the default sensitivity is set to "LOW".

Managed Sensitive Data Types

Global entities

Data Type	Category	Description
PERSON_NAME	Universal	The name of the person.
EMAIL	Personal	The email address.
IP_ADDRESS	Computer	The IP address

Data Type	Category	Description
MAC_ADDRESS	Personal	The MAC address.

US data types

Data Type	Description
BANK_ACCOUNT	The bank account number. Not specific to a country or region, however, only US and Canadian account formats are detected.
CREDIT_CARD	The credit card number.
PHONE_NUMBER	The phone number. Not specific to a country or region, however, only US and Canadian phone numbers are detected at this time.
USA_ATIN	The US Adoption Taxpayer Identification Number issued by the Internal Revenue Service.
USA_CPT_CODE	The CPT Code (US specific).
USA_DEA_NUMBER	The DEA number (US specific).
USA_DRIVING_LICENSE	The driver license number (US specific).
USA_HCPCS_CODE	The HCPCS code (US specific).
USA_HEALTH_INSURANCE_CLAIM_NUMBER	Health Insurance Claim Number (US specific).
USA_ITIN	The ITIN (for US persons or entities).
USA_MEDICARE_BENEFICIARY_IDENTIFIER	Medicare Beneficiary Identifier (US specific).
USA_NATIONAL_DRUG_CODE	The NDC code (US specific).

Data Type	Description
USA_NATIONAL_PROVIDER_IDENTIFIER	The National Provider Identifier number (US specific).
USA_PASSPORT_NUMBER	The passport number (for US persons).
USA_PTIN	The US Preparer Tax Identification Number issued by the Internal Revenue Service.
USA_SSN	The social security number (for US persons).

Argentina data types

Data Type	Description
ARGENTINA_TAX_IDENTIFICATION_NUMBER	Argentina Tax Identification Number. Also known as CUIT or CUIL.

Australian data types

Data Type	Description
AUSTRALIA_BUSINESS_NUMBER	Australia Business Number (ABN). A unique identifier issued by the Australian Business Register (ABR) to identify businesses to the government and community.
AUSTRALIA_COMPANY_NUMBER	Australia Company Number (ACN). Unique identifier issued by the Australian Securities and Investments Commission.
AUSTRALIA_DRIVING_LICENSE	A driver’s license number for Australia.

Data Type	Description
AUSTRALIA_MEDICARE_NUMBER	Australian Medicare Number. Personal identifier issued by the Australian Health Insurance Commission.
AUSTRALIA_PASSPORT_NUMBER	Australian passport number.
AUSTRALIA_TAX_FILE_NUMBER	Australia Tax File Number (TFN). Issued by the Australian Taxation Office (ATO) to taxpayers (individual, company, etc) for tax dealings.

Austria data types

Data Type	Description
AUSTRIA_DRIVING_LICENSE	The driver license number (Austria specific).
AUSTRIA_PASSPORT_NUMBER	The passport number (Austria specific).
AUSTRIA_SSN	The social security number (for Austria persons).
AUSTRIA_TAX_IDENTIFICATION_NUMBER	Tax identification number (Austria specific).
AUSTRIA_VALUE_ADDED_TAX	Value-Added Tax (Austria specific).

Balkans data types

Data Type	Description
BOSNIA_UNIQUE_MASTER_CITIZEN_NUMBER	Unique master citizen number (JMBG) for Bosnia-Herzegovina citizens.
KOSOVO_UNIQUE_MASTER_CITIZEN_NUMBER	Unique master citizen number (JMBG) for Kosovo.

Data Type	Description
MACEDONIA_UNIQUE_MASTER_CITIZEN_NUMBER	Unique master citizen number for Macedonia.
MONTENEGRO_UNIQUE_MASTER_CITIZEN_NUMBER	Unique master citizen number (JMBG) for Montenegro.
SERBIA_UNIQUE_MASTER_CITIZEN_NUMBER	Unique master citizen number (JMBG) for Serbia.
SERBIA_VALUE_ADDED_TAX	Value-Added Tax (Serbia specific).
VOJVODINA_UNIQUE_MASTER_CITIZEN_NUMBER	Unique master citizen number (JMBG) for Vojvodina.

Belgium data types

Data Type	Description
BELGIUM_DRIVING_LICENSE	The driver license number (Belgium specific).
BELGIUM_NATIONAL_IDENTIFICATION_NUMBER	The Belgian National Number (BNN).
BELGIUM_PASSPORT_NUMBER	The passport number (Belgium specific).
BELGIUM_TAX_IDENTIFICATION_NUMBER	Tax identification number (Belgium specific).
BELGIUM_VALUE_ADDED_TAX	Value-Added Tax (Belgium specific).

Brazil data types

Data Type	Description
BRAZIL_BANK_ACCOUNT	The bank account number (Brazil specific).
BRAZIL_NATIONAL_IDENTIFICATION_NUMBER	The national identifier (Brazil specific).

Data Type	Description
BRAZIL_NATIONAL_REGISTRY_OF_LEGAL_ENTITIES_NUMBER	The identification number issued to companies (Brazil specific), also known as the CNPJ.
BRAZIL_NATURAL_PERSON_REGISTRY_NUMBER	Natural Person Registry Number, also known as CPF.

Bulgaria data types

Data Type	Description
BULGARIA_DRIVING_LICENSE	The driver license number (Bulgaria specific).
BULGARIA_UNIFORM_CIVIL_NUMBER	Unified Civil Number (EGN) that serves as a national identification number.
BULGARIA_VALUE_ADDED_TAX	Value-Added Tax (Bulgaria specific).

Canada data types

Data Type	Description
CANADA_DRIVING_LICENSE	The driver license number (Canada specific).
CANADA_GOVERNMENT_IDENTIFICATION_CARD_NUMBER	The national identifier (Canada specific).
CANADA_PASSPORT_NUMBER	The passport number (Canada specific).
CANADA_PERMANENT_RESIDENCE_NUMBER	Permanent residence number (PR Card number).
CANADA_PERSONAL_HEALTH_NUMBER	The unique identifier for healthcare (PHN number).
CANADA_SOCIAL_INSURANCE_NUMBER	The social insurance number (SIN) in Canada.

Chile data types

Data Type	Description
CHILE_DRIVING_LICENSE	The driver license number (Chile specific).
CHILE_NATIONAL_IDENTIFICATION_NUMBER	The Chile national identifier, also known as RUT or RUN.

China, Hong Kong, Macau, and Taiwan data types

Data Type	Description
CHINA_IDENTIFICATION	The China identifier.
CHINA_LICENSE_PLATE_NUMBER	The driver license number (China specific).
CHINA_MAINLAND_TRAVEL_PERMIT_ID_HONG_KONG_MACAU	The Mainland Travel Permit for Hong Kong and Macao Residents.
CHINA_MAINLAND_TRAVEL_PERMIT_ID_TAIWAN	The Mainland Travel Permit for Taiwan Residents issued by Government of the People's Republic of China (PRC).
CHINA_PASSPORT_NUMBER	The passport number (China specific).
CHINA_PHONE_NUMBER	The phone number (China specific).
HONG_KONG_IDENTITY_CARD	The official identity document issued by the Immigration Department of Hong Kong.
MACAU_RESIDENT_IDENTITY_CARD	The Macau Resident Identity Card or BIR is an official identity card issued by the Identification Services Bureau of Macau.
TAIWAN_NATIONAL_IDENTIFICATION_NUMBER	The national identifier (Taiwan specific).
TAIWAN_PASSPORT_NUMBER	The passport number (Taiwan specific).

Colombia data types

Data Type	Description
COLOMBIA_PERSONAL_IDENTIFICATION_NUMBER	Unique identifier assigned to Colombians at birth.
COLOMBIA_TAX_IDENTIFICATION_NUMBER	Tax identification number (Colombia specific).

Croatia data types

Data Type	Description
CROATIA_DRIVING_LICENSE	The driver license number (Croatia specific).
CROATIA_IDENTITY_NUMBER	The national identifier (Croatia specific).
CROATIA_PASSPORT_NUMBER	The passport number (Croatia specific).
CROATIA_PERSONAL_IDENTIFICATION_NUMBER	The personal identifier number (OIB).

Cyprus data types

Data Type	Description
CYPRUS_DRIVING_LICENSE	The driver license number (Cyprus specific).
CYPRUS_NATIONAL_IDENTIFICATION_NUMBER	The Cypriot identity card.
CYPRUS_PASSPORT_NUMBER	The passport number (Cyprus specific).
CYPRUS_TAX_IDENTIFICATION_NUMBER	Tax identification number (Cyprus specific).
CYPRUS_VALUE_ADDED_TAX	Value-Added Tax (Cyprus specific).

Czechia data types

Data Type	Description
CZECHIA_DRIVING_LICENSE	The driver license number (Czechia specific).
CZECHIA_PERSONAL_IDENTIFICATION_NUMBER	The personal identifier number (Czechia specific).
CZECHIA_VALUE_ADDED_TAX	Value-Added Tax (Czechia specific).

Denmark data types

Data Type	Description
DENMARK_DRIVING_LICENSE	The driver license number (Denmark specific).
DENMARK_PERSONAL_IDENTIFICATION_NUMBER	The personal identifier number (Denmark specific).
DENMARK_TAX_IDENTIFICATION_NUMBER	Tax identification number (Denmark specific).
DENMARK_VALUE_ADDED_TAX	Value-Added Tax (Denmark specific).

Estonia data types

Data Type	Description
ESTONIA_DRIVING_LICENSE	The driver license number (Estonia specific).
ESTONIA_PASSPORT_NUMBER	The passport number (Estonia specific).
ESTONIA_PERSONAL_IDENTIFICATION_CODE	The personal identifier number (Estonia specific).
ESTONIA_VALUE_ADDED_TAX	Value-Added Tax (Estonia specific).

Finland data types

Data Type	Description
FINLAND_DRIVING_LICENSE	The driver license number (Finland specific).
FINLAND_HEALTH_INSURANCE_NUMBER	The health insurance number (Finland specific) .
FINLAND_NATIONAL_IDENTIFICATION_NUMBER	The national identifier number (Finland specific).
FINLAND_PASSPORT_NUMBER	The passport number (Finland specific).
FINLAND_VALUE_ADDED_TAX	Value-Added Tax (Finland specific).

France data types

Data Type	Description
FRANCE_BANK_ACCOUNT	The bank account number (France specific).
FRANCE_DRIVING_LICENSE	The driver license number (France specific).
FRANCE_HEALTH_INSURANCE_NUMBER	France health insurance number.
FRANCE_INSEE_CODE	France social security, SSN, or NIR number.
FRANCE_NATIONAL_IDENTIFICATION_NUMBER	France national identifier number (CNI).
FRANCE_PASSPORT_NUMBER	The passport number (France specific).
FRANCE_TAX_IDENTIFICATION_NUMBER	Tax identification number (France specific).
FRANCE_VALUE_ADDED_TAX	Value-Added Tax (France specific).

Germany data types

Data Type	Description
GERMANY_BANK_ACCOUNT	The bank account number (Germany specific).
GERMANY_DRIVING_LICENSE	The driver license number (Germany specific).
GERMANY_PASSPORT_NUMBER	The passport number (Germany specific).
GERMANY_PERSONAL_IDENTIFICATION_NUMBER	The personal identification number (Germany specific).
GERMANY_TAX_IDENTIFICATION_NUMBER	Tax identification number (Germany specific).
GERMANY_VALUE_ADDED_TAX	Value-Added Tax (Germany specific).

Greece data types

Data Type	Description
GREECE_DRIVING_LICENSE	The driver license number (Greece specific).
GREECE_PASSPORT_NUMBER	The passport number (Greece specific).
GREECE_SSN	The social security number (for Greece persons).
GREECE_TAX_IDENTIFICATION_NUMBER	Tax identification number (Greece specific).
GREECE_VALUE_ADDED_TAX	Value-Added Tax (Greece specific).

Hungary data types

Data Type	Description
HUNGARY_DRIVING_LICENSE	The driver license number (Hungary specific).
HUNGARY_PASSPORT_NUMBER	The passport number (Hungary specific).

Data Type	Description
HUNGARY_SSN	The social security number (for Hungary persons).
HUNGARY_TAX_IDENTIFICATION_NUMBER	Tax identification number (Hungary specific).
HUNGARY_VALUE_ADDED_TAX	Value-Added Tax (Hungary specific).

Iceland data types

Data Type	Description
ICELAND_NATIONAL_IDENTIFICATION_NUMBER	The national identifier (Iceland specific).
ICELAND_PASSPORT_NUMBER	The passport number (Iceland specific).
ICELAND_VALUE_ADDED_TAX	Value-Added Tax (Iceland specific).

India data types

Data Type	Description
INDIA_AADHAAR_NUMBER	Aadhaar identification number issued by the Unique Identification Authority of India.
INDIA_PERMANENT_ACCOUNT_NUMBER	India Permanent Account Number (PAN).

Indonesia data types

Data Type	Description
INDONESIA_IDENTITY_CARD_NUMBER	The national identifier (Indonesia specific).

Ireland data types

Data Type	Description
IRELAND_DRIVING_LICENSE	The driver license number (Ireland specific).
IRELAND_PASSPORT_NUMBER	The passport number (Ireland specific).
IRELAND_PERSONAL_PUBLIC_SERVICE_NUMBER	Ireland personal public service number (PPS).
IRELAND_TAX_IDENTIFICATION_NUMBER	Tax identification number (Ireland specific).
IRELAND_VALUE_ADDED_TAX	Value-Added Tax (Ireland specific).

Israel data types

Data Type	Description
ISRAEL_IDENTIFICATION_NUMBER	The national identifier (Israel specific).

Italy data types

Data Type	Description
ITALY_BANK_ACCOUNT	The bank account number (Italy specific).
ITALY_DRIVING_LICENSE	The driver license number (Italy specific).
ITALY_FISCAL_CODE	The identifier number, also known as the Italian Codice Fiscale.
ITALY_PASSPORT_NUMBER	The passport number (Italy specific).
ITALY_VALUE_ADDED_TAX	Value-Added Tax (Italy specific).

Japan data types

Data Type	Description
JAPAN_BANK_ACCOUNT	Japan bank account.
JAPAN_DRIVING_LICENSE	A driver's license number for Japan.
JAPAN_MY_NUMBER	The unique identifier for Japan citizens or corporations used for tax administration, social security administration, and disaster response
JAPAN_PASSPORT_NUMBER	Japan passort number.

Korea data types

Data Type	Description
KOREA_PASSPORT_NUMBER	The passport number (Korea specific).
KOREA_RESIDENCE_REGISTRATION_NUMBER_FOR_CITIZENS	Korea residence registrant number for residents.
KOREA_RESIDENCE_REGISTRATION_NUMBER_FOR_FOREIGNERS	Korea residence registrant number for foreigners.

Latvia data types

Data Type	Description
LATVIA_DRIVING_LICENSE	The driver license number (Latvia specific).
LATVIA_PASSPORT_NUMBER	The passport number (Latvia specific).
LATVIA_PERSONAL_IDENTIFICATION_NUMBER	The personal identifier number (Latvia specific).
LATVIA_VALUE_ADDED_TAX	Value-Added Tax (Latvia specific).

Liechtenstein data types

Data Type	Description
LIECHTENSTEIN_NATIONAL_IDENTIFICATION_NUMBER	The national identifier (Liechtenstein specific).
LIECHTENSTEIN_PASSPORT_NUMBER	The passport number (Liechtenstein specific).
LIECHTENSTEIN_TAX_IDENTIFICATION_NUMBER	Tax identification number (Liechtenstein specific).

Lithuania data types

Data Type	Description
LITHUANIA_DRIVING_LICENSE	The driver license number (Lithuania specific).
LITHUANIA_PERSONAL_IDENTIFICATION_NUMBER	The personal identifier number (Lithuania specific).
LITHUANIA_TAX_IDENTIFICATION_NUMBER	Tax identification number (Lithuania specific).
LITHUANIA_VALUE_ADDED_TAX	Value-Added Tax (Lithuania specific).

Luxembourg data types

Data Type	Description
LUXEMBOURG_DRIVING_LICENSE	The driver license number (Luxembourg specific).
LUXEMBOURG_NATIONAL_INDIVIDUAL_NUMBER	The national identifier (Luxembourg specific).
LUXEMBOURG_PASSPORT_NUMBER	The passport number (Luxembourg specific).

Data Type	Description
LUXEMBOURG_TAX_IDENTIFICATION_NUMBER	Tax identification number (Luxembourg specific).
LUXEMBOURG_VALUE_ADDED_TAX	Value-Added Tax (Luxembourg specific).

Malaysia data types

Data Type	Description
MALAYSIA_MYKAD_NUMBER	The national identifier (Malaysia specific).
MALAYSIA_PASSPORT_NUMBER	The passport number (Malaysia specific).

Malta data types

Data Type	Description
MALTA_DRIVING_LICENSE	The driver license number (Malta specific).
MALTA_NATIONAL_IDENTIFICATION_NUMBER	The national identifier (Malta specific).
MALTA_TAX_IDENTIFICATION_NUMBER	Tax identification number (Malta specific).
MALTA_VALUE_ADDED_TAX	Value-Added Tax (Malta specific).

Mexico data types

Data Type	Description
MEXICO_CLABE_NUMBER	Mexico CLABE (Clave Bancaria Estandarizada) bank number).
MEXICO_DRIVING_LICENSE	The driver license number (Mexico specific).
MEXICO_PASSPORT_NUMBER	The passport number (Mexico specific).

Data Type	Description
MEXICO_TAX_IDENTIFICATION_NUMBER	Tax identification number (Mexico specific).
MEXICO_UNIQUE_POPULATION_REGISTRY_CODE	The Clave Única de Registro de Población (CURP) unique identity code for Mexico.

Netherlands data types

Data Type	Description
NETHERLANDS_CITIZEN_SERVICE_NUMBER	Netherlands citizen number (BSN, burgerservicenummer).
NETHERLANDS_DRIVING_LICENSE	The driver license number (Netherlands specific).
NETHERLANDS_PASSPORT_NUMBER	The passport number (Netherlands specific).
NETHERLANDS_TAX_IDENTIFICATION_NUMBER	Tax identification number (Netherlands specific).
NETHERLANDS_VALUE_ADDED_TAX	Value-Added Tax (Netherlands specific).
NETHERLANDS_BANK_ACCOUNT	The bank account number (Netherlands specific).

New Zealand data types

Data Type	Description
NEW_ZEALAND_DRIVING_LICENSE	The driver license number (New Zealand specific).
NEW_ZEALAND_NATIONAL_HEALTH_INDEX_NUMBER	New Zealand national health index number.

Data Type	Description
NEW_ZEALAND_TAX_IDENTIFICATION_NUMBER	Tax identification number, also known as inland revenue number (New Zealand specific).

Norway data types

Data Type	Description
NORWAY_BIRTH_NUMBER	Norwegian national identity number.
NORWAY_DRIVING_LICENSE	The driver license number (Norway specific).
NORWAY_HEALTH_INSURANCE_NUMBER	Norway health insurance number.
NORWAY_NATIONAL_IDENTIFICATION_NUMBER	The national identifier number (Norway specific).
NORWAY_VALUE_ADDED_TAX	Value-Added Tax (Norway specific).

Philippines data types

Data Type	Description
PHILIPPINES_DRIVING_LICENSE	The driver license number (Philippines specific) .
PHILIPPINES_PASSPORT_NUMBER	The passport number (Philippines specific).

Poland data types

Data Type	Description
POLAND_DRIVING_LICENSE	The driver license number (Poland specific).
POLAND_IDENTIFICATION_NUMBER	The Poland identifier.

Data Type	Description
POLAND_PASSPORT_NUMBER	The passport number (Poland specific).
POLAND_REGON_NUMBER	The REGON identifier number, also known as the Statistical Identification Number.
POLAND_SSN	The social security number (for Poland persons).
POLAND_TAX_IDENTIFICATION_NUMBER	Tax identification number (Poland specific).
POLAND_VALUE_ADDED_TAX	Value-Added Tax (Poland specific).

Portugal data types

Data Type	Description
PORTUGAL_DRIVING_LICENSE	The driver license number (Portugal specific).
PORTUGAL_NATIONAL_IDENTIFICATION_NUMBER	The national identifier number (Portugal specific).
PORTUGAL_PASSPORT_NUMBER	The passport number (Portugal specific).
PORTUGAL_TAX_IDENTIFICATION_NUMBER	Tax identification number (Portugal specific).
PORTUGAL_VALUE_ADDED_TAX	Value-Added Tax (Portugal specific).

Romania data types

Data Type	Description
ROMANIA_DRIVING_LICENSE	The driver license number (Romania specific).
ROMANIA_NUMERICAL_PERSONAL_CODE	The personal identifier number (Romania specific).

Data Type	Description
ROMANIA_PASSPORT_NUMBER	The passport number (Romania specific).
ROMANIA_VALUE_ADDED_TAX	Value-Added Tax (Romania specific).

Singapore data types

Data Type	Description
SINGAPORE_DRIVING_LICENSE	The driver license number (Singapore specific).
SINGAPORE_NATIONAL_REGISTRY_IDENTIFICATION_NUMBER	The national registration identity card for Singapore.
SINGAPORE_PASSPORT_NUMBER	The passport number (Singapore specific).
SINGAPORE_UNIQUE_ENTITY_NUMBER	The Unique Entity Number for Singapore.

Slovakia data types

Data Type	Description
SLOVAKIA_DRIVING_LICENSE	The driver license number (Slovakia specific).
SLOVAKIA_NATIONAL_IDENTIFICATION_NUMBER	The national identifier number (Slovakia specific).
SLOVAKIA_PASSPORT_NUMBER	The passport number (Slovakia specific).
SLOVAKIA_VALUE_ADDED_TAX	Value-Added Tax (Slovakia specific).

Slovenia data types

Data Type	Description
SLOVENIA_DRIVING_LICENSE	The driver license number (Slovenia specific).

Data Type	Description
SLOVENIA_PASSPORT_NUMBER	The passport number (Slovenia specific).
SLOVENIA_TAX_IDENTIFICATION_NUMBER	Tax identification number (Slovenia specific).
SLOVENIA_UNIQUE_MASTER_CITIZEN_NUMBER	Unique master citizen number (JMBG) for Slovenia citizens.
SLOVENIA_VALUE_ADDED_TAX	Value-Added Tax (Slovenia specific).

South Africa data types

Data Type	Description
SOUTH_AFRICA_PERSONAL_IDENTIFICATION_NUMBER	The personal identifier number (South Sfrica specific).

Spain data types

Data Type	Description
SPAIN_BANK_ACCOUNT	The bank account number (Spain specific).
SPAIN_DNI	The national identity card (Documento Nacional de Identidad) of Spain.
SPAIN_DRIVING_LICENSE	The driver license number (Spain specific).
SPAIN_NIE	The foreigner identity number (Spain specific), also known as the NIE.
SPAIN_NIF	Tax identification number (Spain specific), also known as the NIF.
SPAIN_PASSPORT_NUMBER	The passport number (Spain specific).
SPAIN_SSN	The social security number (for Spain persons).

Data Type	Description
SPAIN_VALUE_ADDED_TAX	Value-Added Tax (Spain specific).

Sri Lanka data types

Data Type	Description
SRI_LANKA_NATIONAL_IDENTIFICATION_NUMBER	The national identifier (Sri Lanka specific).

Sweden data types

Data Type	Description
SWEDEN_DRIVING_LICENSE	The driver license number (Sweden specific).
SWEDEN_PASSPORT_NUMBER	The passport number (Sweden specific).
SWEDEN_PERSONAL_IDENTIFICATION_NUMBER	The national identifier number (Sweden specific).
SWEDEN_TAX_IDENTIFICATION_NUMBER	Sweden tax identification number (personnummer).
SWEDEN_VALUE_ADDED_TAX	Value-Added Tax (Sweden specific).

Switzerland data types

Data Type	Description
SWITZERLAND_AHV	The social security number for Swiss persons (AHV).
SWITZERLAND_HEALTH_INSURANCE_NUMBER	Swiss health insurance number.

Data Type	Description
SWITZERLAND_PASSPORT_NUMBER	The passport number (Switzerland specific).
SWITZERLAND_VALUE_ADDED_TAX	Value-Added Tax (Switzerland specific).

Thailand data types

Data Type	Description
THAILAND_PASSPORT_NUMBER	The passport number (Thailand specific).
THAILAND_PERSONAL_IDENTIFICATION_NUMBER	The personal identifier number (Thailand specific).

Turkey data types

Data Type	Description
TURKEY_NATIONAL_IDENTIFICATION_NUMBER	The national identifier number (Turkey specific).
TURKEY_PASSPORT_NUMBER	The passport number (Turkey specific).
TURKEY_VALUE_ADDED_TAX	Value-Added Tax (Turkey specific).

Ukraine data types

Data Type	Description
UKRAINE_INDIVIDUAL_IDENTIFICATION_NUMBER	The unique identifier (Ukraine specific).
UKRAINE_PASSPORT_NUMBER_DOMESTIC	The domestic passport number (Ukraine specific).

Data Type	Description
UKRAINE_PASSPORT_NUMBER_INTERNATIONAL	The international passport number (Ukraine specific).

United Arab Emirates (UAE) data types

Data Type	Description
UNITED_ARAB_EMIRATES_PERSONAL_NUMBER	The personal identifier number (UAE specific).

UK data types

Data Type	Description
UK_BANK_ACCOUNT	United Kingdom (UK) bank account.
UK_BANK_SORT_CODE	United Kingdom (UK) bank sort code. Sort codes are bank codes used to route money transfers between banks within their respective countries via their respective clearance organizations.
UK_DRIVING_LICENSE	The driver's license number for the United Kingdom of Great Britain and Northern Ireland (UK specific)
UK_ELECTORAL_ROLL_NUMBER	The Electoral Roll Number (ERN) is the identification number issued to an individual for UK election registration. The format of this number is specified by the UK Government Standards of the UK Cabinet Office.
UK_NATIONAL_HEALTH_SERVICE_NUMBER	The National Health Service (NHS) number is the unique number allocated to a registere

Data Type	Description
	d user of public health services in the United Kingdom.
UK_NATIONAL_INSURANCE_NUMBER	The National Insurance number (NINO) is a number used in the United Kingdom (UK) to identify an individual for the national insurance program or social security system. The number is sometimes referred to as NI No or NINO.
UK_PASSPORT_NUMBER	United Kingdom (UK) passport number.
UK_UNIQUE_TAXPAYER_REFERENCE_NUMBER	The United Kingdom (UK) Unique Taxpayer Reference (UTR) number. An identifier used by the UK government to manage the taxation system.
UK_VALUE_ADDED_TAX	VAT is a consumption tax that is borne by the end consumer. VAT is paid for each transaction in the manufacturing and distribution process. For the United Kingdom, the VAT number is issued by the VAT office for the region in which the business is established.
UK_PHONE_NUMBER	United Kingdom (UK) phone number.

Venezuela data types

Data Type	Description
VENEZUELA_DRIVING_LICENSE	The driver license number (Venezuela specific).
VENEZUELA_NATIONAL_IDENTIFICATION_NUMBER	The national identifier number (Venezuela specific).
VENEZUELA_VALUE_ADDED_TAX	Value-Added Tax (Venezuela specific).

Using fine-grained sensitive data detection

> **Note**
>
> Fine-grained actions is only available in AWS Glue 3.0 and 4.0. This includes the AWS Glue Studio experience. The persistent audit log changes are also not available in 2.0.
> All AWS Glue Studio 3.0 and 4.0 visual jobs will have a script created that automatically uses fine-grained actions APIs.

The Detect Sensitive Data transform provides the ability to detect, mask, or remove entities that you define, or are pre-defined by AWS Glue. Fine-grained actions further allows you to apply a specific action per entity. Additional benefits include:

- Improved performance as actions are being applied as soon data is detected.
- The option to include or exclude specific columns.
- The ability to use partial masking. This allows you to mask detected sensitive data entities partially, rather than masking the entire string. Both simple params with offsets and regex are supported.

The following are code snippets of sensitive data detection APIs and fine-grained actions used in the sample jobs referenced in the next section.

Detect API – fine-grained actions use the new `detectionParameters` parameter:

```
def detect(
    frame: DynamicFrame,
    detectionParameters: JsonOptions,
    outputColumnName: String = "DetectedEntities",
    detectionSensitivity: String = "LOW"
): DynamicFrame = {}
```

Using Sensitive Data Detection APIs with fine-grained actions

Sensitive data detection APIs using **detect** analyzes the data given, determines if the rows or columns are Sensitive Data Entity Types, and will run actions specified by the user for each Entity type.

Using the detect API with fine-grained actions

Use the **detect** API and specify the `outputColumnName` and `detectionParameters`.

```
object GlueApp {
  def main(sysArgs: Array[String]) {

    val spark: SparkContext = new SparkContext()
    val glueContext: GlueContext = new GlueContext(spark)

    // @params: [JOB_NAME]
    val args = GlueArgParser.getResolvedOptions(sysArgs, Seq("JOB_NAME").toArray)
    Job.init(args("JOB_NAME"), glueContext, args.asJava)

    // Script generated for node S3 bucket. Creates DataFrame from data stored in
S3.
    val S3bucket_node1 =
glueContext.getSourceWithFormat(formatOptions=JsonOptions("""{"quoteChar":
"\"", "withHeader": true, "separator": ",", "optimizePerformance": false}"""),
connectionType="s3", format="csv", options=JsonOptions("""{"paths":
["s3://189657479688-ddevansh-pii-test-bucket/tiny_pii.csv"], "recurse": true}"""),
transformationContext="S3bucket_node1").getDynamicFrame()

    // Script generated for node Detect Sensitive Data. Will run detect API for the
DataFrame
    // detectionParameter contains information on which EntityType are being
detected
    // and what actions are being applied to them when detected.
    val DetectSensitiveData_node2 = EntityDetector.detect(
        frame = S3bucket_node1,
        detectionParameters = JsonOptions(
         """
            {
                "PHONE_NUMBER": [
                    {
                        "action": "PARTIAL_REDACT",
                        "actionOptions": {
                            "numLeftCharsToExclude": "3",
                            "numRightCharsToExclude": "4",
                            "redactChar": "#"
                        },
                        "sourceColumnsToExclude": [ "Passport No", "DL NO#" ]
                    }
                    ],
                    "USA_PASSPORT_NUMBER": [
                        {
                            "action": "SHA256_HASH",
                            "sourceColumns": [ "Passport No" ]
                        }
                    ],
                    "USA_DRIVING_LICENSE": [
                        {
                            "action": "REDACT",
                            "actionOptions": {
                                "redactText": "USA_DL"
                            },
                            "sourceColumns": [ "DL NO#" ]
                        }
                    ]

                }
            """
            ),
            outputColumnName = "DetectedEntities"
        )

        // Script generated for node S3 bucket. Store Results of detect to S3 location
        val S3bucket_node3 = glueContext.getSinkWithFormat(connectionType="s3",
 options=JsonOptions("""{"path": "s3://amzn-s3-demo-bucket/test-output/",
 "partitionKeys": []}"""), transformationContext="S3bucket_node3",
 format="json").writeDynamicFrame(DetectSensitiveData_node2)

        Job.commit()
      }
```

The above script will create a DataFrame from a location in Amazon S3 and then it will run the `detect` API. Since the `detect` API requires the field `detectionParameters` (a map of the entity name to a list all of the action settings to be used for that entity) is represented by AWS Glue's `JsonOptions` object, it will also allow us to extend the functionality of the API.

For each action specified per entity, enter a list of all column names to which to apply the entity/action combination. This allows you to customize the entities to detect for every column in your dataset and skip entities that you know are not in a specific column. This also allows your jobs to be more performant by not performing unnecessary detection calls those entities and allows you to perform actions unique to each column and entity combination.

Taking a closer look at the `detectionParameters`, there are three entity types in the sample job. These are `Phone Number`, `USA_PASSPORT_NUMBER`, and `USA_DRIVING_LICENSE`. For each of these entity types AWS Glue will run different actions which are either `PARTIAL_REDACT`, `SHA256_HASH`, `REDACT`, and `DETECT`. Each of the Entity Types also have `sourceColumns` to apply to and/or `sourceColumnsToExclude` if detected.

> **Note**
>
> Only one edit-in-place action (`PARTIAL_REDACT`, `SHA256_HASH`, or `REDACT`) can be used per column but the `DETECT` action can be used with any of these actions.

The `detectionParameters` field has the below layout:

```
    ENTITY_NAME -> List[Actions]
    {
     "ENTITY_NAME": [{
      Action, // required
      ColumnSpecs,
      ActionOptionsMap
        }],
        "ENTITY_NAME2": [{
      ...
        }]
    }
```

The types of `actions` and `actionOptions` are listed below:

```
DETECT
{
    # Required
    "action": "DETECT",
    # Optional, depending on action chosen
    "actionOptions": {
        // There are no actionOptions for DETECT
    },
    # 1 of below required, both can also used
    "sourceColumns": [
        "COL_1", "COL_2", ..., "COL_N"
    ],
    "sourceColumnsToExclude": [
        "COL_5"
    ]
}

SHA256_HASH
{
    # Required
    "action": "SHA256_HASH",
    # Required or optional, depending on action chosen
    "actionOptions": {
        // There are no actionOptions for SHA256_HASH
    },

    # 1 of below required, both can also used
    "sourceColumns": [
        "COL_1", "COL_2", ..., "COL_N"
    ],
    "sourceColumnsToExclude": [
        "COL_5"
    ]
}

REDACT
{
    # Required
    "action": "REDACT",
    # Required or optional, depending on action chosen
    "actionOptions": {
        // The text that is being replaced
        "redactText": "USA_DL"
    },

    # 1 of below required, both can also used
    "sourceColumns": [
        "COL_1", "COL_2", ..., "COL_N"
    ],
    "sourceColumnsToExclude": [
        "COL_5"
    ]
```

```
}

PARTIAL_REDACT
{
    # Required
    "action": "PARTIAL_REDACT",
    # Required or optional, depending on action chosen
    "actionOptions": {
        // number of characters to not redact from the left side
        "numLeftCharsToExclude": "3",
        // number of characters to not redact from the right side
        "numRightCharsToExclude": "4",
        // the partial redact will be made with this redacted character
        "redactChar": "#",
        // regex pattern for partial redaction
        "matchPattern": "[0-9]"
    },

    # 1 of below required, both can also used
    "sourceColumns": [
        "COL_1", "COL_2", ..., "COL_N"
    ],
    "sourceColumnsToExclude": [
        "COL_5"
    ]
}
```

Once the script runs, results are output to the given Amazon S3 location. You can view your data in Amazon S3 but with the selected entity types being sensitized based on the selected action. In the case, we would have a rows that would have that looked like this:

```
{
    "Name": "Colby Schuster",
    "Address": "39041 Antonietta Vista, South Rodgerside, Nebraska 24151",
    "Car Owned": "Fiat",
    "Email": "Kitty46@gmail.com",
    "Company": "O'Reilly Group",
    "Job Title": "Dynamic Functionality Facilitator",
    "ITIN": "991-22-2906",
    "Username": "Cassandre.Kub43",
    "SSN": "914-22-2906",
    "DOB": "2020-08-27",
    "Phone Number": "1-2#######1718",
    "Bank Account No": "69741187",
    "Credit Card Number": "6441-6289-6867-2162-2711",
    "Passport No": "94f311e93a623c72ccb6fc46cf5f5b0265ccb42c517498a0f27fd4c43b47111e",
    "DL NO#": "USA_DL"
}
```

In the above script, the `Phone Number` was partially redacted with `#`. The `Passport No` was changed into a SHA256 hash. The `DL NO#` was detected as a USA driver license number and was redacted to “USA_DL” just like it was stated in the `detectionParameters`.

> **Note**
>
> The classifyColumns API is not available for use with fine-grained actions due to the nature of the API. This API performs column sampling (adjustable by the user but has default values) to perform detection more quickly. Fine-grained actions require iterating over every value for this reason.

Persistent Audit Log

A new feature introduced with fine-grained actions (but also available when using the normal APIs) is the presence of a persistent audit log. Currently, running the detect API adds an additional column (defaults to `DetectedEntities` but customizable through the `outputColumnName`) parameter with PII detection metadata. This now has an “actionUsed” metadata key, which is one of `DETECT`, `PARTIAL_REDACT`, `SHA256_HASH`, `REDACT`.

```
"DetectedEntities": {
    "Credit Card Number": [
        {
            "entityType": "CREDIT_CARD",
            "actionUsed": "DETECT",
            "start": 0,
            "end": 19
        }
    ],
    "Phone Number": [
        {
            "entityType": "PHONE_NUMBER",
            "actionUsed": "REDACT",
            "start": 0,
            "end": 14
        }
    ]
}
```

Even customers using APIs without fine-grained actions such as `detect(entityTypesToDetect, outputColumnName)` will see this persistent audit log in the resulting dataframe.

Customers using APIs with fine-grained actions will see all of the actions, regardless of if they are redacted or not. Example:

```
+---------------------+----------------
+---------------------------------------------------------------------------------------------
+
| Credit Card Number  |  Phone Number  |
                                             DetectedEntities
                                                                  |
+---------------------+----------------
+---------------------------------------------------------------------------------------------
+
| 622126741306XXXX    | +12#####7890   | {"Credit Card Number":
[{"entityType":"CREDIT_CARD","actionUsed":"PARTIAL_REDACT","start":0,"end":16}],"Phone
 Number":
[{"entityType":"PHONE_NUMBER","actionUsed":"PARTIAL_REDACT","start":0,"end":12}]}} |
| 6221 2674 1306 XXXX | +12#######7890 | {"Credit Card Number":
[{"entityType":"CREDIT_CARD","actionUsed":"PARTIAL_REDACT","start":0,"end":19}],"Phone
 Number":
[{"entityType":"PHONE_NUMBER","actionUsed":"PARTIAL_REDACT","start":0,"end":14}]}} |
| 6221-2674-1306-XXXX | 22#######7890  | {"Credit Card Number":
[{"entityType":"CREDIT_CARD","actionUsed":"PARTIAL_REDACT","start":0,"end":19}],"Phone
 Number":
[{"entityType":"PHONE_NUMBER","actionUsed":"PARTIAL_REDACT","start":0,"end":14}]}} |
+---------------------+----------------
+---------------------------------------------------------------------------------------------
+
```

If you do not want to see the **DetectedEntities** column, you can simply drop the additional column in a custom script.

AWS Glue Visual Job API

AWS Glue provides an API that allows customers to create data integration jobs using the AWS Glue API from a JSON object that represents a visual step workflow. Customers can then use the visual editor in AWS Glue Studio to work with these jobs.

For more information on Visual Job API data types, see Visual Job API.

Topics

- API design and CRUD APIs
- Getting started
- Visual job limitations

API design and CRUD APIs

The CreateJob and UpdateJob APIs now support an additional optional parameter, codeGenConfigurationNodes. Providing a non-empty JSON structure for this field will result in the DAG being registered in AWS Glue Studio for the created job and the associated code being generated. A null value or empty string for this field on job create will be ignored.

Updates to the codeGenConfigurationNodes field will be done through the UpdateJob AWS Glue API in a similar way as CreateJob. The entire field should be specified in UpdateJob where the DAG has been changed as desired. A null value provided will be ignored and no update to the DAG would be performed. An empty structure or string will cause the codeGenConfigurationNodes to be set as empty and any previous DAG removed. The GetJob API will return a DAG if one exists. The DeleteJob API will also delete any associated DAG.

Getting started

To create a job, use the CreateJob action. The `CreateJob` request input will have an additional field ‘codeGenConfigurationNodes’ where you can get specify the DAG object in JSON.

Things to keep in mind:

- The ‘codeGenConfigurationNodes’ field is a map of nodeId to node.
- Each node begins with a key identifying what kind of node it is.
- There can only be one key specified since a node can only be of one type.
- The input field contains the parent nodes of the current node.

The following is a JSON representation of a **CreateJob** input.

```
{
  "node-1": {
    "S3CatalogSource": {
      "Table": "csvFormattedTable",
      "PartitionPredicate": "",
      "Name": "S3 bucket",
      "AdditionalOptions": {},
      "Database": "myDatabase"
    }
  },
  "node-3": {
    "S3DirectTarget": {
      "Inputs": ["node-2"],
      "PartitionKeys": [],
      "Compression": "none",
      "Format": "json",
      "SchemaChangePolicy": { "EnableUpdateCatalog": false },
      "Path": "",
      "Name": "S3 bucket"
    }
  },
  "node-2": {
    "ApplyMapping": {
      "Inputs": ["node-1"],
      "Name": "ApplyMapping",
      "Mapping": [
        {
          "FromType": "long",
          "ToType": "long",
          "Dropped": false,
          "ToKey": "myheader1",
          "FromPath": ["myheader1"]
        },
        {
          "FromType": "long",
          "ToType": "long",
          "Dropped": false,
          "ToKey": "myheader2",
          "FromPath": ["myheader2"]
        },
        {
          "FromType": "long",
          "ToType": "long",
          "Dropped": false,
          "ToKey": "myheader3",
          "FromPath": ["myheader3"]
        }
      ]
    }
  }
}
```

Updating and getting jobs

Since *UpdateJob* will also have a ‘codeGenConfigurationNodes’ field, the input format will be the same. See UpdateJob Action.

The *GetJob* action will return a ‘codeGenConfigurationNodes’ field in the same format as well. See GetJob Action.

Visual job limitations

Since the ‘codeGenConfigurationNodes’ parameter has been added to existing APIs, any limitations in those APIs will be inherited. In addition, the codeGenConfigurationNodes and some nodes will be limited in size. See Job Structure for more information.

Programming Ray scripts

AWS Glue makes it easy to write and run Ray scripts. This section describes the supported Ray capabilities that are available in AWS Glue for Ray. You program Ray scripts in Python.

Your custom script must be compatible with the version of Ray that's defined by the `Runtime` field in your job definition. For more information about `Runtime` in the Jobs API, see the section called “Jobs”. For information about each runtime environment, see the section called “Supported Ray runtime environments”.

Topics

- Tutorial: Writing an ETL script in AWS Glue for Ray
- Using Ray Core and Ray Data in AWS Glue for Ray
- Providing files and Python libraries to Ray jobs

- Connecting to data in Ray jobs

Tutorial: Writing an ETL script in AWS Glue for Ray

Ray gives you the ability to write and scale distributed tasks natively in Python. AWS Glue for Ray offers serverless Ray environments that you can access from both jobs and interactive sessions (Ray interactive sessions are in preview). The AWS Glue job system provides a consistent way to manage and run your tasks—on a schedule, from a trigger, or from the AWS Glue console.

Combining these AWS Glue tools creates a powerful toolchain that you can use for extract, transform, and load (ETL) workloads, a popular use case for AWS Glue. In this tutorial, you will learn the basics of putting together this solution.

We also support using AWS Glue for Spark for your ETL workloads. For a tutorial on writing a AWS Glue for Spark script, see the section called “Tutorial: Writing a Spark script”. For more information about available engines, see the section called “AWS Glue for Spark and AWS Glue for Ray”. Ray is capable of addressing many different kinds of tasks in analytics, machine learning (ML), and application development.

In this tutorial, you will extract, transform, and load a CSV dataset that is hosted in Amazon Simple Storage Service (Amazon S3). You will begin with the New York City Taxi and Limousine Commission (TLC) Trip Record Data Dataset, which is stored in a public Amazon S3 bucket. For more information about this dataset, see the Registry of Open Data on AWS.

You will transform your data with predefined transforms that are available in the Ray Data library. Ray Data is a dataset preparation library designed by Ray and included by default in AWS Glue for Ray environments. For more information about libraries included by default, see the section called “Modules provided with Ray jobs”. You will then write your transformed data to an Amazon S3 bucket that you control.

Prerequisites – For this tutorial, you need an AWS account with access to AWS Glue and Amazon S3.

Step 1: Create a bucket in Amazon S3 to hold your output data

You will need an Amazon S3 bucket that you control to serve as a sink for data created in this tutorial. You can create this bucket with the following procedure.

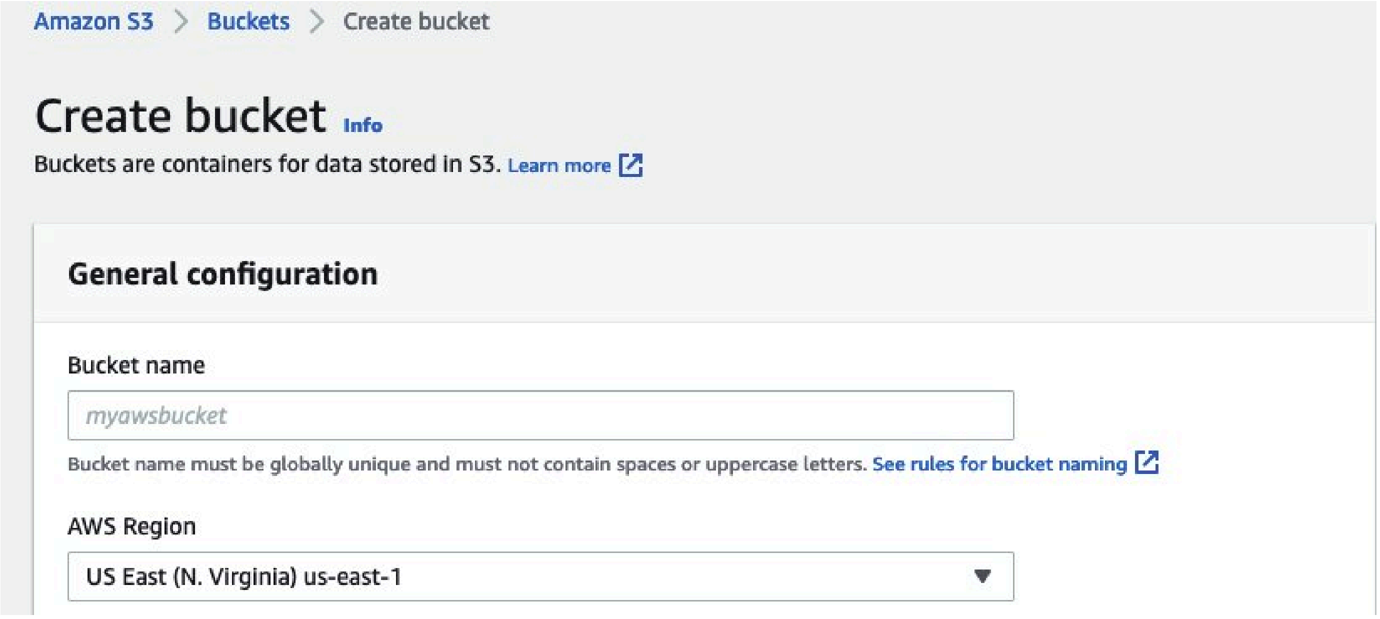
Note

If you want to write your data to an existing bucket that you control, you can skip this step. Take note of *yourBucketName*, the existing bucket's name, to use in later steps.

To create a bucket for your Ray job output

- Create a bucket by following the steps in [Creating a bucket](#) in the *Amazon S3 User Guide*.
  - When choosing a bucket name, take note of *yourBucketName*, which you will refer to in later steps.
  - For other configuration, the suggested settings provided in the Amazon S3 console should work fine in this tutorial.

  As an example, the bucket creation dialog box might look like this in the Amazon S3 console.



Amazon S3 > Buckets > Create bucket

Create bucket Info

Buckets are containers for data stored in S3. Learn more

General configuration

Bucket name

myawsbucket

Bucket name must be globally unique and must not contain spaces or uppercase letters. See rules for bucket naming

AWS Region

US East (N. Virginia) us-east-1

Step 2: Create an IAM role and policy for your Ray job

Your job will need an AWS Identity and Access Management (IAM) role with the following:

- Permissions granted by the `AWSGlueServiceRole` managed policy. These are the basic permissions that are necessary to run an AWS Glue job.

- `Read` access level permissions for the `nyc-tlc/*` Amazon S3 resource.
- `Write` access level permissions for the *`yourBucketName`*`/*` Amazon S3 resource.
- A trust relationship that allows the `glue.amazonaws.com` principal to assume the role.

You can create this role with the following procedure.

To create an IAM role for your AWS Glue for Ray job

> **Note**
>
> You can create an IAM role by following many different procedures. For more information or options about how to provision IAM resources, see the AWS Identity and Access Management documentation.

1. Create a policy that defines the previously outlined Amazon S3 permissions by following the steps in Creating IAM policies (console) with the visual editor in the *IAM User Guide*.
   - When selecting a service, choose Amazon S3.
   - When selecting permissions for your policy, attach the following sets of actions for the following resources (mentioned previously):
     - Read access level permissions for the `nyc-tlc/*` Amazon S3 resource.
     - Write access level permissions for the *`yourBucketName`*`/*` Amazon S3 resource.
   - When selecting the policy name, take note of *`YourPolicyName`*, which you will refer to in a later step.
2. Create a role for your AWS Glue for Ray job by following the steps in Creating a role for an AWS service (console) in the *IAM User Guide*.
   - When selecting a trusted AWS service entity, choose `Glue`. This will automatically populate the necessary trust relationship for your job.
   - When selecting policies for the permissions policy, attach the following policies:
     - `AWSGlueServiceRole`
     - *`YourPolicyName`*
   - When selecting the role name, take note of *`YourRoleName`*, which you will refer to in later steps.

Step 3: Create and run an AWS Glue for Ray job

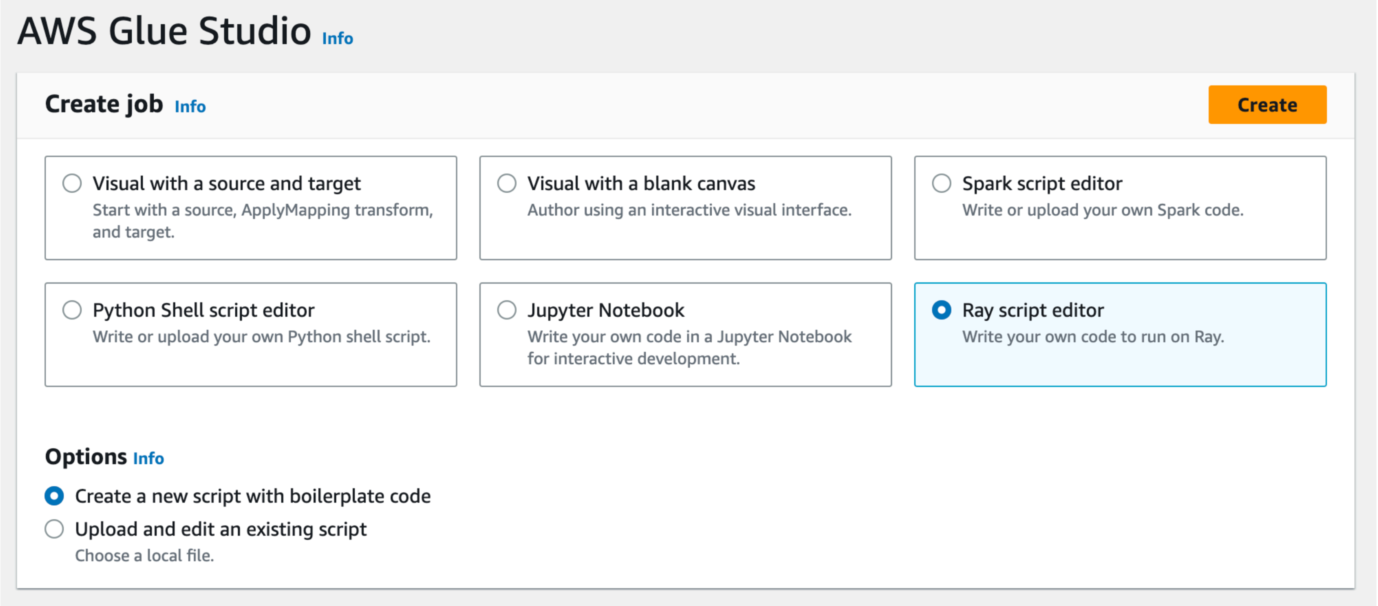
In this step, you create an AWS Glue job using the AWS Management Console, provide it with a sample script, and run the job. When you create a job, it creates a place in the console for you to store, configure, and edit your Ray script. For more information about creating jobs, see the section called "Signing in to the console".

In this tutorial, we address the following ETL scenario: you would like to read the January 2022 records from the New York City TLC Trip Record dataset, add a new column (`tip_rate`) to the dataset by combining data in existing columns, then remove a number of columns that aren't relevant to your current analysis, and then you would like to write the results to *`yourBucketName`*. The following Ray script performs these steps:

```
import ray
import pandas
from ray import data

ray.init('auto')

ds = ray.data.read_csv("s3://nyc-tlc/opendata_repo/opendata_webconvert/yellow/
yellow_tripdata_2022-01.csv")

# Add the given new column to the dataset and show the sample record after adding a new
 column
ds = ds.add_column( "tip_rate", lambda df: df["tip_amount"] / df["total_amount"])

# Dropping few columns from the underlying Dataset
ds = ds.drop_columns(["payment_type", "fare_amount", "extra", "tolls_amount",
 "improvement_surcharge"])

ds.write_parquet("s3://yourBucketName/ray/tutorial/output/")
```

To create and run an AWS Glue for Ray job

1. In the AWS Management Console, navigate to the AWS Glue landing page.
2. In the side navigation pane, choose **ETL Jobs**.
3. In **Create job**, choose **Ray script editor**, and then choose **Create**, as in the following illustration.



AWS Glue Studio Info

Create job Info

○ Visual with a source and target
Start with a source, ApplyMapping transform, and target.

○ Visual with a blank canvas
Author using an interactive visual interface.

○ Spark script editor
Write or upload your own Spark code.

○ Python Shell script editor
Write or upload your own Python shell script.

○ Jupyter Notebook
Write your own code in a Jupyter Notebook for interactive development.

● Ray script editor
Write your own code to run on Ray.

Options Info

● Create a new script with boilerplate code

○ Upload and edit an existing script
Choose a local file.

4. Paste the full text of the script into the **Script** pane, and replace any existing text.
5. Navigate to **Job details** and set the **IAM Role** property to *`YourRoleName`*.
6. Choose **Save**, and then choose **Run**.

Step 4: Inspect your output

After running your AWS Glue job, you should validate that the output matches the expectations of this scenario. You can do so with the following procedure.

To validate whether your Ray job ran successfully

1. On the job details page, navigate to **Runs**.
2. After a few minutes, you should see a run with a **Run status** of **Succeeded**.
3. Navigate to the Amazon S3 console at https://console.aws.amazon.com/s3/ and inspect *`yourBucketName`*. You should see files written to your output bucket.
4. Read the Parquet files and verify their contents. You can do this with your existing tools. If you don't have a process for validating Parquet files, you can do this in the AWS Glue console with an AWS Glue interactive session, using either Spark or Ray (in preview).

   In an interactive session, you have access to Ray Data, Spark, or pandas libraries, which are provided by default (based on your choice of engine). To verify your file contents, you can use common inspection methods that are available in those libraries—methods like `count`, `schema`, and `show`. For more information about interactive sessions in the console, see Using notebooks with AWS Glue Studio and AWS Glue.

Because you have confirmed that files have been written to the bucket, you can say with relative certainty that if your output has problems, they are not related to IAM configuration. Configure your session with *`yourRoleName`* to have access to the relevant files.

If you don't see the expected outcomes, examine the troubleshooting content in this guide to identify and remediate the source of the error. You can find the troubleshooting content in the *Troubleshooting AWS Glue* chapter. For specific errors that are related to Ray jobs, see the section called “Troubleshooting Ray errors” in the troubleshooting chapter.

Next steps

You have now seen and performed an ETL process using AWS Glue for Ray from end to end. You can use the following resources to understand what tools AWS Glue for Ray provides to transform and interpret your data at scale.

- For more information about Ray's task model, see the section called “Using Ray Core and Ray Data in AWS Glue for Ray”. For more experience in using Ray tasks, follow the examples in the Ray Core documentation. See Ray Core: Ray Tutorials and Examples (2.4.0) in the Ray documentation.
- For guidance about available data management libraries in AWS Glue for Ray, see the section called “Connecting to data”. For more experience using Ray Data to transform and write datasets, follow the examples in the Ray Data documentation. See Ray Data: Examples (2.4.0).
- For more information about configuring AWS Glue for Ray jobs, see the section called “Working with Ray jobs”.
- For more information about writing AWS Glue for Ray scripts, continue reading the documentation in this section.

Using Ray Core and Ray Data in AWS Glue for Ray

Ray is a framework for scaling up Python scripts by distributing work across a cluster. You can use Ray as a solution to many sorts of problems, so Ray provides libraries to optimize certain tasks. In AWS Glue, we focus on using Ray to transform large datasets. AWS Glue offers support for Ray Data and parts of Ray Core to facilitate this task.

What is Ray Core?

The first step of building a distributed application is identifying and defining work that can be performed concurrently. Ray Core contains the parts of Ray that you use to define tasks that can be performed concurrently. Ray provides reference and quick start information that you can use to learn the tools they provide. For more information, see What is Ray Core? and Ray Core Quick Start. For more information about effectively defining concurrent tasks in Ray, see Tips for first-time users.

Ray tasks and actors

In AWS Glue for Ray documentation, we might refer to *tasks* and *actors*, which are core concepts in Ray.

Ray uses Python functions and classes as the building blocks of a distributed computing system. Much like when Python functions and variables become "methods" and "attributes" when used in a class, functions become "tasks" and classes become "actors" when they're used in Ray to send code to workers. You can identify functions and classes that might be used by Ray by the `@ray.remote` annotation.

Tasks and actors are configurable, they have a lifecycle, and they take up compute resources throughout their life. Code that throws errors can be traced back to a task or actor when you're finding the root cause of problems. Thus, these terms might come up when you're learning how to configure, monitor, or debug AWS Glue for Ray jobs.

To begin learning how to effectively use tasks and actors to build a distributed application, see Key Concepts in the Ray docs.

Ray Core in AWS Glue for Ray

AWS Glue for Ray environments manage cluster formation and scaling, as well as collecting and visualizing logs. Because we manage these concerns, we consequently limit access to and support for the APIs in Ray Core that would be used to address these concerns in an open-source cluster.

In the managed `Ray2.4` runtime environment, we do not support:

- Ray Core CLI
- Ray State CLI
- `ray.util.metrics` Prometheus metric utility methods:
  - Counter

  - Gauge
  - Histogram
- Other debugging tools:
  - ray.util.pdb.set_trace
  - ray.util.inspect_serializability
  - ray.timeline

What is Ray Data?

When you're connecting to data sources and destinations, handling datasets, and initiating common transforms, Ray Data is a straightforward methodology for using Ray to solve problems transforming Ray datasets. For more information about using Ray Data, see Ray Datasets: Distributed Data Preprocessing.

You can use Ray Data or other tools to access your data. For more information on accessing your data in Ray, see the section called “Connecting to data”.

Ray Data in AWS Glue for Ray

Ray Data is supported and provided by default in the managed `Ray2.4` runtime environment. For more information about provided modules, see the section called “Modules provided with Ray jobs”.

Providing files and Python libraries to Ray jobs

This section provides information that you need for using Python libraries with AWS Glue Ray jobs. You can use certain common libraries included by default in all Ray jobs. You can also provide your own Python libraries to your Ray job.

Modules provided with Ray jobs

You can perform data integration workflows in a Ray job with the following provided packages. These packages are available by default in Ray jobs.

AWS Glue version 4.0

In AWS Glue 4.0, the Ray (`Ray2.4` runtime) environment provides the following packages:

- boto3 == 1.26.133
- ray == 2.4.0
- pyarrow == 11.0.0
- pandas == 1.5.3
- numpy == 1.24.3
- fsspec == 2023.4.0

This list includes all packages that would be installed with `ray[data] == 2.4.0`. Ray Data is supported out of box.

Providing files to your Ray job

You can provide files to your Ray job with the `--working-dir` parameter. Provide this parameter with a path to a .zip file hosted on Amazon S3. Within the .zip file, your files must be contained in a single top-level directory. No other files should be at the top level.

Your files will be distributed to each Ray node before your script begins to run. Consider how this might impact the disk space that's available to each Ray node. Available disk space is determined by the WorkerType set in the job configuration. If you want to provide your job data at scale, this mechanism is not the right solution. For more information on providing data to your job, see the section called “Connecting to data”.

Your files will be accessible as if the directory was provided to Ray through the `working_dir` parameter. For example, to read a file named `sample.txt` in your .zip file's top-level directory, you could call:

```
@ray.remote
def do_work():
    f = open("sample.txt", "r")
    print(f.read())
```

For more information about `working_dir`, see the Ray documentation. This feature behaves similarly to Ray's native capabilities.

Additional Python modules for Ray jobs

Additional modules from PyPI

Ray jobs use the Python Package Installer (pip3) to install additional modules to be used by a Ray script. You can use the `--pip-install` parameter with a list of comma-separated Python modules to add a new module or change the version of an existing module.

For example, to update or add a new `scikit-learn` module, use the following key-value pair:

```
"--pip-install", "scikit-learn==0.21.3"
```

If you have custom modules or custom patches, you can distribute your own libraries from Amazon S3 with the `--s3-py-modules` parameter. Before uploading your distribution, it might need to be repackaged and rebuilt. Follow the guidelines in in the section called "Including Python code in Ray jobs".

Custom distributions from Amazon S3

Custom distributions should adhere to Ray packaging guidelines for dependencies. You can find out how to build these distributions in the following section. For more information about how Ray sets up dependencies, see Environment Dependencies in the Ray documentation.

To include a custom distributable after assessing its contents, upload your distributable to a bucket available to the job's IAM role. Specify the Amazon S3 path to a Python zip archive in your parameter configuration. If you're providing multiple distributables, separate them by comma. For example:

```
"--s3-py-modules", "s3://s3bucket/pythonPackage.zip"
```

Limitations

Ray jobs do not support compiling native code in the job environment. You can be limited by this if your Python dependencies transitively depend on native, compiled code. Ray jobs can run provided binaries, but they must be compiled for Linux on ARM64. This means you might be able to use the contents of `aarch64manylinux` wheels. You can provide your native dependencies in a compiled form by repackaging a wheel to Ray standards. Typically, this means removing `dist-info` folders so that there is only one folder at the root at the archive.

You cannot upgrade the version of `ray` or `ray[data]` using this parameter. In order to use a new version of Ray, you will need to change the runtime field on your job, after we have released support for it. For more information about supported Ray versions, see the section called "AWS Glue versions".

Including Python code in Ray jobs

The Python Software Foundation offers standardized behaviors for packaging Python files for use across different runtimes. Ray introduces limitations to packaging standards that you should be aware of. AWS Glue does not specify packaging standards beyond those specified to Ray. The following instructions provide standard guidance on packaging simple Python packages.

Package your files in a `.zip` archive. A directory should be at the root of the archive. **There should be no other files at the root level of the archive, or this may lead to unexpected behavior.** The root directory is the package, and its name is used to refer to your Python code when importing it.

If you provide a distribution in this form to a Ray job with `--s3-py-modules`, you will be able to import Python code from your package in your Ray script.

Your package can provide a single Python module with some Python files, or you can package together many modules. When repackaging dependencies, such as libraries from PyPI, **check for hidden files and metadata directories** inside of those packages.

⚠ **Warning**

Certain OS behaviors make make it difficult to properly follow these packaging instructions.

- OSX may add hidden files such as `__MACOSX` to your zip file at the top level.
- Windows may add your files to a folder inside the zip automatically, unintentionally creating a nested folder.

The following procedures assume you are interacting with your files in Amazon Linux 2 or a similar OS that provides a distribution of the Info-ZIP `zip` and `zipinfo` utilities. We recommend using these tools to prevent unexpected behaviors.

To package Python files for use in Ray

1. Create a temporary directory with your package name, then confirm your working directory is its parent directory. You can do this with the following commands:

   ```
   cd parent_directory
   mkdir temp_dir
   ```

2. Copy your files into the temporary directory, then confirm your directory structure. The contents of this directory will be directly accessed as your Python module. You can do this with the following command:

   ```
   ls -AR temp_dir
   # my_file_1.py
   # my_file_2.py
   ```

3. Compress your temporary folder using zip. You can do this with the following commands:

   ```
   zip -r zip_file.zip temp_dir
   ```

4. Confirm your file is properly packaged. *`zip_file`*`.zip` should now be found in your working directory. You can inspect it with the following command:

   ```
   zipinfo -1 zip_file.zip
   # temp_dir/
   # temp_dir/my_file_1.py
   # temp_dir/my_file_2.py
   ```

To repackage a Python package for use in Ray.

1. Create a temporary directory with your package name, then confirm your working directory is its parent directory. You can do this with the following commands:

   ```
   cd parent_directory
   mkdir temp_dir
   ```

2. Decompress your package and copy the contents into your temporary directory. Remove files related to your previous packaging standard, leaving only the contents of the module. Confirm your file structure looks correct with the following command:

   ```
   ls -AR temp_dir
   # my_module
   # my_module/__init__.py
   # my_module/my_file_1.py
   # my_module/my_submodule/__init__.py
   # my_module/my_submodule/my_file_2.py
   ```

```
# my_module/my_submodule/my_file_3.py
```

3. Compress your temporary folder using zip. You can do this with the following commands:

```
zip -r zip_file.zip temp_dir
```

4. Confirm your file is properly packaged. *`zip_file`*`.zip` should now be found in your working directory. You can inspect it with the following command:

```
zipinfo -1 zip_file.zip
# temp_dir/my_module/
# temp_dir/my_module/__init__.py
# temp_dir/my_module/my_file_1.py
# temp_dir/my_module/my_submodule/
# temp_dir/my_module/my_submodule/__init__.py
# temp_dir/my_module/my_submodule/my_file_2.py
# temp_dir/my_module/my_submodule/my_file_3.py
```

Connecting to data in Ray jobs

AWS Glue Ray jobs can use a broad array of Python packages that are designed for you to quickly integrate data. We provide a minimal set of dependencies in order to not clutter your environment. For more information about what is included by default, see the section called “Modules provided with Ray jobs”.

> **Note**
>
> AWS Glue extract, transform, and load (ETL) provides the DynamicFrame abstraction to streamline ETL workflows where you resolve schema differences between rows in your dataset. AWS Glue ETL provides additional features—job bookmarks and grouping input files. We don't currently provide corresponding features in Ray jobs.
> AWS Glue for Spark provides direct support for connecting to certain data formats, sources and sinks. In Ray, AWS SDK for pandas and current third-party libraries substantively cover that need. You will need to consult those libraries to understand what capabilities are available.

AWS Glue for Ray integration with Amazon VPC is not currently available. Resources in Amazon VPC will not be accessible without a public route. For more information about using AWS Glue with Amazon VPC, see the section called “Configuring interface VPC endpoints (AWS PrivateLink) for AWS Glue ”.

Common libraries for working with data in Ray

Ray Data – Ray Data provides methods to handle common data formats, sources and sinks. For more information about supported formats and sources in Ray Data, see Input/Output in the Ray Data documentation. Ray Data is an opinionated library, rather than a general-purpose library, for handling datasets.

Ray provides certain guidance around use cases where Ray Data might be the best solution for your job. For more information, see Ray use cases in the Ray documentation.

AWS SDK for pandas (awswrangler) – AWS SDK for pandas is an AWS product that delivers clean, tested solutions for reading from and writing to AWS services when your transformations manage data with pandas DataFrames. For more information about supported formats and sources in the AWS SDK for pandas, see the API Reference in the AWS SDK for pandas documentation.

For examples of how to read and write data with the AWS SDK for pandas, see Quick Start in the AWS SDK for pandas documentation. The AWS SDK for pandas doesn't provide transforms for your data. It only provides support for reading and writing from sources.

Modin – Modin is a Python library that implements common pandas operations in a distributable way. For more information about Modin, see the Modin documentation. Modin itself doesn't provide support for reading and writing from sources. It provides distributed implementations of common transforms. Modin is supported by the AWS SDK for pandas.

When you run Modin and the AWS SDK for pandas together in a Ray environment, you can perform common ETL tasks with performant results. For more information about using Modin with the AWS SDK for pandas, see At scale in the AWS SDK for pandas documentation.

Other frameworks – For more information about frameworks that Ray supports, see The Ray Ecosystem in the Ray documentation. We don't provide support for other frameworks in AWS Glue for Ray.

Connecting to data through the Data Catalog

Managing your data through the Data Catalog in conjunction with Ray jobs is supported with the AWS SDK for pandas. For more information, see Glue Catalog on the AWS SDK for pandas website.

Using this service with an AWS SDK

AWS software development kits (SDKs) are available for many popular programming languages. Each SDK provides an API, code examples, and documentation that make it easier for developers to build applications in their preferred language.

SDK documentation	Code examples
AWS SDK for C++	AWS SDK for C++ code examples
AWS CLI	AWS CLI code examples
AWS SDK for Go	AWS SDK for Go code examples
AWS SDK for Java	AWS SDK for Java code examples
AWS SDK for JavaScript	AWS SDK for JavaScript code examples
AWS SDK for Kotlin	AWS SDK for Kotlin code examples
AWS SDK for .NET	AWS SDK for .NET code examples
AWS SDK for PHP	AWS SDK for PHP code examples
AWS Tools for PowerShell	AWS Tools for PowerShell code examples
AWS SDK for Python (Boto3)	AWS SDK for Python (Boto3) code examples
AWS SDK for Ruby	AWS SDK for Ruby code examples
AWS SDK for Rust	AWS SDK for Rust code examples
AWS SDK for SAP ABAP	AWS SDK for SAP ABAP code examples
AWS SDK for Swift	AWS SDK for Swift code examples

For examples specific to this service, see AWS Glue API code examples using AWS SDKs.

Example availability

Can't find what you need? Request a code example by using the **Provide feedback** link at the bottom of this page.

AWS Glue API

This section describes data types and primitives used by AWS Glue SDKs and Tools. There are three general ways to interact with AWS Glue programmatically outside of the AWS Management Console, each with its own documentation:

- Language SDK libraries allow you to access AWS resources from common programming languages. Find more information at Tools to Build on AWS.
- The AWS CLI allows you to access AWS resources from the command line. Find more information at AWS CLI Command Reference.
- AWS CloudFormation allows you to define a set of AWS resources to be provisioned together consistently. Find more information at AWS CloudFormation: AWS Glue resource type reference.

This section documents shared primitives independently of these SDKs and Tools. Tools use the AWS Glue Web API Reference to communicate with AWS.

Contents

- Security APIs in AWS Glue
  - Data types
  - DataCatalogEncryptionSettings structure
  - EncryptionAtRest structure
  - ConnectionPasswordEncryption structure
  - EncryptionConfiguration structure
  - S3Encryption structure
  - CloudWatchEncryption structure
  - JobBookmarksEncryption structure
  - SecurityConfiguration structure
  - GluePolicy structure
  - DataQualityEncryption structure
  - Operations
  - GetDataCatalogEncryptionSettings action (Python: get_data_catalog_encryption_settings)
  - PutDataCatalogEncryptionSettings action (Python: put_data_catalog_encryption_settings)
  - PutResourcePolicy action (Python: put_resource_policy)

  - GetResourcePolicy action (Python: get_resource_policy)
  - DeleteResourcePolicy action (Python: delete_resource_policy)
  - CreateSecurityConfiguration action (Python: create_security_configuration)
  - DeleteSecurityConfiguration action (Python: delete_security_configuration)
  - GetSecurityConfiguration action (Python: get_security_configuration)
  - GetSecurityConfigurations action (Python: get_security_configurations)
  - GetResourcePolicies action (Python: get_resource_policies)
- Catalog objects API
  - Catalogs API
    - Data types
    - Catalog structure
    - CatalogInput structure
    - TargetRedshiftCatalog structure
    - CatalogProperties structure
    - CatalogPropertiesOutput structure
    - DataLakeAccessProperties structure
    - IcebergOptimizationProperties structure
    - DataLakeAccessPropertiesOutput structure
    - IcebergOptimizationPropertiesOutput structure
    - FederatedCatalog structure
    - Operations
    - CreateCatalog action (Python: create_catalog)
    - UpdateCatalog action (Python: update_catalog)
    - DeleteCatalog action (Python: delete_catalog)
    - GetCatalog action (Python: get_catalog)
    - GetCatalogs action (Python: get_catalogs)
  - Database API
    - Data types
    - Database structure
    - DatabaseInput structure

    - PrincipalPermissions structure
    - DataLakePrincipal structure
    - DatabaseIdentifier structure
    - FederatedDatabase structure
    - Operations
    - CreateDatabase action (Python: create_database)
    - UpdateDatabase action (Python: update_database)
    - DeleteDatabase action (Python: delete_database)
    - GetDatabase action (Python: get_database)
    - GetDatabases action (Python: get_databases)
- Table API
    - Data types
    - Table structure
    - TableInput structure
    - FederatedTable structure
    - Column structure
    - StorageDescriptor structure
    - SchemaReference structure
    - SerDeInfo structure
    - Order structure
    - SkewedInfo structure
    - TableVersion structure
    - TableError structure
    - TableVersionError structure
    - SortCriterion structure
    - TableIdentifier structure
    - KeySchemaElement structure
    - PartitionIndex structure
    - PartitionIndexDescriptor structure
    - BackfillError structure

- IcebergInput structure
- OpenTableFormatInput structure
- ViewDefinition structure
- ViewDefinitionInput structure
- ViewRepresentation structure
- ViewRepresentationInput structure
- UpdateOpenTableFormatInput structure
- UpdateIcebergInput structure
- CreateIcebergTableInput structure
- UpdateIcebergTableInput structure
- IcebergSortOrder structure
- IcebergSortField structure
- IcebergPartitionSpec structure
- IcebergPartitionField structure
- IcebergSchema structure
- IcebergStructField structure
- IcebergTableUpdate structure
- Operations
- CreateTable action (Python: create_table)
- UpdateTable action (Python: update_table)
- DeleteTable action (Python: delete_table)
- BatchDeleteTable action (Python: batch_delete_table)
- GetTable action (Python: get_table)
- GetTables action (Python: get_tables)
- GetTableVersion action (Python: get_table_version)
- GetTableVersions action (Python: get_table_versions)
- DeleteTableVersion action (Python: delete_table_version)
- BatchDeleteTableVersion action (Python: batch_delete_table_version)
- SearchTables action (Python: search_tables)
- GetPartitionIndexes action (Python: get_partition_indexes)

  - CreatePartitionIndex action (Python: create_partition_index)
  - DeletePartitionIndex action (Python: delete_partition_index)
  - GetColumnStatisticsForTable action (Python: get_column_statistics_for_table)
  - UpdateColumnStatisticsForTable action (Python: update_column_statistics_for_table)
  - DeleteColumnStatisticsForTable action (Python: delete_column_statistics_for_table)
- Partition API
  - Data types
  - Partition structure
  - PartitionInput structure
  - PartitionSpecWithSharedStorageDescriptor structure
  - PartitionListComposingSpec structure
  - PartitionSpecProxy structure
  - PartitionValueList structure
  - Segment structure
  - PartitionError structure
  - BatchUpdatePartitionFailureEntry structure
  - BatchUpdatePartitionRequestEntry structure
  - StorageDescriptor structure
  - SchemaReference structure
  - SerDeInfo structure
  - SkewedInfo structure
  - Operations
  - CreatePartition action (Python: create_partition)
  - BatchCreatePartition action (Python: batch_create_partition)
  - UpdatePartition action (Python: update_partition)
  - DeletePartition action (Python: delete_partition)
  - BatchDeletePartition action (Python: batch_delete_partition)
  - GetPartition action (Python: get_partition)
  - GetPartitions action (Python: get_partitions)
  - BatchGetPartition action (Python: batch_get_partition)

    - BatchUpdatePartition action (Python: batch_update_partition)
    - GetColumnStatisticsForPartition action (Python: get_column_statistics_for_partition)
    - UpdateColumnStatisticsForPartition action (Python: update_column_statistics_for_partition)
    - DeleteColumnStatisticsForPartition action (Python: delete_column_statistics_for_partition)
- Connections API
    - Connection API
        - Data types
        - Connection structure
        - ConnectionInput structure
        - TestConnectionInput structure
        - PhysicalConnectionRequirements structure
        - GetConnectionsFilter structure
        - AuthenticationConfiguration structure
        - AuthenticationConfigurationInput structure
        - OAuth2Properties structure
        - OAuth2PropertiesInput structure
        - OAuth2ClientApplication structure
        - AuthorizationCodeProperties structure
        - BasicAuthenticationCredentials structure
        - OAuth2Credentials structure
        - Operations
        - CreateConnection action (Python: create_connection)
        - DeleteConnection action (Python: delete_connection)
        - GetConnection action (Python: get_connection)
        - GetConnections action (Python: get_connections)
        - UpdateConnection action (Python: update_connection)
        - TestConnection action (Python: test_connection)
        - BatchDeleteConnection action (Python: batch_delete_connection)
    - Connection Types API

        - Connection management APIs
        - DescribeConnectionType action (Python: describe_connection_type)
        - ListConnectionTypes action (Python: list_connection_types)
        - ConnectionTypeBrief structure
        - ConnectionTypeVariant structure
        - datatypes
        - Validation structure
        - AuthConfiguration structure
        - Capabilities structure
        - Property structure
        - AllowedValue structure
        - ComputeEnvironmentConfiguration structure
    - Connection Metadata and Preview API
        - Data types
        - Entity structure
        - Field structure
        - Operations
        - ListEntities action (Python: list_entities)
        - DescribeEntity action (Python: describe_entity)
        - GetEntityRecords action (Python: get_entity_records)
- User-defined Function API
    - Data types
    - UserDefinedFunction structure
    - UserDefinedFunctionInput structure
    - Operations
    - CreateUserDefinedFunction action (Python: create_user_defined_function)
    - UpdateUserDefinedFunction action (Python: update_user_defined_function)
    - DeleteUserDefinedFunction action (Python: delete_user_defined_function)
    - GetUserDefinedFunction action (Python: get_user_defined_function)

    - GetUserDefinedFunctions action (Python: get_user_defined_functions)

  - Importing an Athena catalog to AWS Glue
    - Data types
    - CatalogImportStatus structure
    - Operations
    - ImportCatalogToGlue action (Python: import_catalog_to_glue)
    - GetCatalogImportStatus action (Python: get_catalog_import_status)
- Table optimizer API
  - Data types
  - TableOptimizer structure
  - TableOptimizerConfiguration structure
  - TableOptimizerVpcConfiguration structure
  - CompactionConfiguration structure
  - IcebergCompactionConfiguration structure
  - TableOptimizerRun structure
  - BatchGetTableOptimizerEntry structure
  - BatchTableOptimizer structure
  - BatchGetTableOptimizerError structure
  - RetentionConfiguration structure
  - IcebergRetentionConfiguration structure
  - OrphanFileDeletionConfiguration structure
  - IcebergOrphanFileDeletionConfiguration structure
  - CompactionMetrics structure
  - RetentionMetrics structure
  - OrphanFileDeletionMetrics structure
  - IcebergCompactionMetrics structure
  - IcebergRetentionMetrics structure
  - IcebergOrphanFileDeletionMetrics structure
  - RunMetrics structure
  - Operations
  - GetTableOptimizer action (Python: get_table_optimizer)

- BatchGetTableOptimizer action (Python: batch_get_table_optimizer)
- ListTableOptimizerRuns action (Python: list_table_optimizer_runs)
- CreateTableOptimizer action (Python: create_table_optimizer)
- DeleteTableOptimizer action (Python: delete_table_optimizer)
- UpdateTableOptimizer action (Python: update_table_optimizer)

- Crawlers and classifiers API
  - Classifier API
    - Data types
    - Classifier structure
    - GrokClassifier structure
    - XMLClassifier structure
    - JsonClassifier structure
    - CsvClassifier structure
    - CreateGrokClassifierRequest structure
    - UpdateGrokClassifierRequest structure
    - CreateXMLClassifierRequest structure
    - UpdateXMLClassifierRequest structure
    - CreateJsonClassifierRequest structure
    - UpdateJsonClassifierRequest structure
    - CreateCsvClassifierRequest structure
    - UpdateCsvClassifierRequest structure
    - Operations
    - CreateClassifier action (Python: create_classifier)
    - DeleteClassifier action (Python: delete_classifier)
    - GetClassifier action (Python: get_classifier)
    - GetClassifiers action (Python: get_classifiers)
    - UpdateClassifier action (Python: update_classifier)
  - Crawler API
    - Data types
    - Crawler structure

- Schedule structure
- CrawlerTargets structure
- S3Target structure
- S3DeltaCatalogTarget structure
- S3DeltaDirectTarget structure
- JdbcTarget structure
- MongoDBTarget structure
- DynamoDBTarget structure
- DeltaTarget structure
- IcebergTarget structure
- HudiTarget structure
- CatalogTarget structure
- CrawlerMetrics structure
- CrawlerHistory structure
- CrawlsFilter structure
- SchemaChangePolicy structure
- LastCrawlInfo structure
- RecrawlPolicy structure
- LineageConfiguration structure
- LakeFormationConfiguration structure
- Operations
- CreateCrawler action (Python: create_crawler)
- DeleteCrawler action (Python: delete_crawler)
- GetCrawler action (Python: get_crawler)
- GetCrawlers action (Python: get_crawlers)
- GetCrawlerMetrics action (Python: get_crawler_metrics)
- UpdateCrawler action (Python: update_crawler)
- StartCrawler action (Python: start_crawler)
- StopCrawler action (Python: stop_crawler)
- BatchGetCrawlers action (Python: batch_get_crawlers)

  - ListCrawlers action (Python: list_crawlers)
  - ListCrawls action (Python: list_crawls)
- Column statistics API
  - Data types
  - ColumnStatisticsTaskRun structure
  - ColumnStatisticsTaskSettings structure
  - ExecutionAttempt structure
  - Operations
  - StartColumnStatisticsTaskRun action (Python: start_column_statistics_task_run)
  - GetColumnStatisticsTaskRun action (Python: get_column_statistics_task_run)
  - GetColumnStatisticsTaskRuns action (Python: get_column_statistics_task_runs)
  - ListColumnStatisticsTaskRuns action (Python: list_column_statistics_task_runs)
  - StopColumnStatisticsTaskRun action (Python: stop_column_statistics_task_run)
  - CreateColumnStatisticsTaskSettings action (Python: create_column_statistics_task_settings)
  - UpdateColumnStatisticsTaskSettings action (Python: update_column_statistics_task_settings)
  - GetColumnStatisticsTaskSettings action (Python: get_column_statistics_task_settings)
  - DeleteColumnStatisticsTaskSettings action (Python: delete_column_statistics_task_settings)
  - StartColumnStatisticsTaskRunSchedule action (Python: start_column_statistics_task_run_schedule)
  - StopColumnStatisticsTaskRunSchedule action (Python: stop_column_statistics_task_run_schedule)
  - Exceptions
  - ColumnStatisticsTaskRunningException structure
  - ColumnStatisticsTaskNotRunningException structure
  - ColumnStatisticsTaskStoppingException structure
  - ColumnStatisticsTaskAutoConcurrencyLimitException structure
  - InvalidCatalogSettingException structure

- Crawler scheduler API

    - Data types
    - Schedule structure
    - Operations
    - UpdateCrawlerSchedule action (Python: update_crawler_schedule)
    - StartCrawlerSchedule action (Python: start_crawler_schedule)
    - StopCrawlerSchedule action (Python: stop_crawler_schedule)
- Autogenerating ETL Scripts API
  - Data types
  - CodeGenNode structure
  - CodeGenNodeArg structure
  - CodeGenEdge structure
  - Location structure
  - CatalogEntry structure
  - MappingEntry structure
  - Operations
  - CreateScript action (Python: create_script)
  - GetDataflowGraph action (Python: get_dataflow_graph)
  - GetMapping action (Python: get_mapping)
  - GetPlan action (Python: get_plan)
- Visual job API
  - Data types
  - CodeGenConfigurationNode structure
  - JDBCConnectorOptions structure
  - StreamingDataPreviewOptions structure
  - AthenaConnectorSource structure
  - JDBCConnectorSource structure
  - SparkConnectorSource structure
  - CatalogSource structure
  - MySQLCatalogSource structure

  - PostgreSQLCatalogSource structure

- OracleSQLCatalogSource structure
- MicrosoftSQLServerCatalogSource structure
- CatalogKinesisSource structure
- DirectKinesisSource structure
- KinesisStreamingSourceOptions structure
- CatalogKafkaSource structure
- DirectKafkaSource structure
- KafkaStreamingSourceOptions structure
- RedshiftSource structure
- AmazonRedshiftSource structure
- AmazonRedshiftNodeData structure
- AmazonRedshiftAdvancedOption structure
- Option structure
- S3CatalogSource structure
- S3SourceAdditionalOptions structure
- S3CsvSource structure
- DirectJDBCSource structure
- S3DirectSourceAdditionalOptions structure
- S3JsonSource structure
- S3ParquetSource structure
- S3DeltaSource structure
- S3CatalogDeltaSource structure
- CatalogDeltaSource structure
- S3HudiSource structure
- S3CatalogHudiSource structure
- S3ExcelSource structure
- CatalogHudiSource structure
- DynamoDBCatalogSource structure
- RelationalCatalogSource structure
- JDBCConnectorTarget structure

- SparkConnectorTarget structure
- BasicCatalogTarget structure
- MySQLCatalogTarget structure
- PostgreSQLCatalogTarget structure
- OracleSQLCatalogTarget structure
- MicrosoftSQLServerCatalogTarget structure
- RedshiftTarget structure
- AmazonRedshiftTarget structure
- UpsertRedshiftTargetOptions structure
- S3CatalogTarget structure
- S3GlueParquetTarget structure
- CatalogSchemaChangePolicy structure
- S3DirectTarget structure
- S3HudiCatalogTarget structure
- S3HudiDirectTarget structure
- S3DeltaCatalogTarget structure
- S3DeltaDirectTarget structure
- S3HyperDirectTarget structure
- S3IcebergDirectTarget structure
- DirectSchemaChangePolicy structure
- ApplyMapping structure
- Mapping structure
- SelectFields structure
- DropFields structure
- RenameField structure
- Spigot structure
- Join structure
- JoinColumn structure
- SplitFields structure
- SelectFromCollection structure

- FillMissingValues structure
- Filter structure
- FilterExpression structure
- FilterValue structure
- CustomCode structure
- SparkSQL structure
- SqlAlias structure
- DropNullFields structure
- NullCheckBoxList structure
- NullValueField structure
- Datatype structure
- Merge structure
- Union structure
- PIIDetection structure
- Aggregate structure
- DropDuplicates structure
- GovernedCatalogTarget structure
- GovernedCatalogSource structure
- AggregateOperation structure
- GlueSchema structure
- GlueStudioSchemaColumn structure
- GlueStudioColumn structure
- DynamicTransform structure
- TransformConfigParameter structure
- EvaluateDataQuality structure
- DQResultsPublishingOptions structure
- DQStopJobOnFailureOptions structure
- EvaluateDataQualityMultiFrame structure
- Recipe structure
- RecipeReference structure

  - SnowflakeNodeData structure
  - SnowflakeSource structure
  - SnowflakeTarget structure
  - ConnectorDataSource structure
  - ConnectorDataTarget structure
  - RecipeStep structure
  - RecipeAction structure
  - ConditionExpression structure
  - S3CatalogIcebergSource structure
  - CatalogIcebergSource structure
  - S3IcebergCatalogTarget structure
  - DynamoDBELTConnectorSource structure
  - DDBELTConnectionOptions structure
  - DDBELTCatalogAdditionalOptions structure
  - Route structure
  - GroupFilters structure
  - AutoDataQuality structure
- Jobs API
  - Jobs
    - Data types
    - Job structure
    - ExecutionProperty structure
    - NotificationProperty structure
    - JobCommand structure
    - ConnectionsList structure
    - JobUpdate structure
    - SourceControlDetails structure
    - Operations
    - CreateJob action (Python: create_job)
    - UpdateJob action (Python: update_job)

  - GetJob action (Python: get_job)
  - GetJobs action (Python: get_jobs)
  - DeleteJob action (Python: delete_job)
  - ListJobs action (Python: list_jobs)
  - BatchGetJobs action (Python: batch_get_jobs)
- Job runs
  - Data types
  - JobRun structure
  - Predecessor structure
  - JobBookmarkEntry structure
  - BatchStopJobRunSuccessfulSubmission structure
  - BatchStopJobRunError structure
  - NotificationProperty structure
  - Operations
  - StartJobRun action (Python: start_job_run)
  - BatchStopJobRun action (Python: batch_stop_job_run)
  - GetJobRun action (Python: get_job_run)
  - GetJobRuns action (Python: get_job_runs)
  - GetJobBookmark action (Python: get_job_bookmark)
  - GetJobBookmarks action (Python: get_job_bookmarks)
  - ResetJobBookmark action (Python: reset_job_bookmark)
- Triggers
  - Data types
  - Trigger structure
  - TriggerUpdate structure
  - Predicate structure
  - Condition structure
  - Action structure
  - EventBatchingCondition structure
  - Operations

    - CreateTrigger action (Python: create_trigger)
    - StartTrigger action (Python: start_trigger)
    - GetTrigger action (Python: get_trigger)
    - GetTriggers action (Python: get_triggers)
    - UpdateTrigger action (Python: update_trigger)
    - StopTrigger action (Python: stop_trigger)
    - DeleteTrigger action (Python: delete_trigger)
    - ListTriggers action (Python: list_triggers)
    - BatchGetTriggers action (Python: batch_get_triggers)
- Integration APIs in AWS Glue
  - Data types
  - Integration structure
  - IntegrationConfig structure
  - IntegrationPartition structure
  - IntegrationError structure
  - IntegrationFilter structure
  - InboundIntegration structure
  - SourceProcessingProperties structure
  - TargetProcessingProperties structure
  - SourceTableConfig structure
  - TargetTableConfig structure
  - Operations
  - CreateIntegration action (Python: create_integration)
  - ModifyIntegration action (Python: modify_integration)
  - DescribeIntegrations action (Python: describe_integrations)
  - DeleteIntegration action (Python: delete_integration)
  - DescribeInboundIntegrations action (Python: describe_inbound_integrations)
  - CreateIntegrationTableProperties action (Python: create_integration_table_properties)
  - UpdateIntegrationTableProperties action (Python: update_integration_table_properties)
  - GetIntegrationTableProperties action (Python: get_integration_table_properties)

  - DeleteIntegrationTableProperties action (Python: delete_integration_table_properties)
  - CreateIntegrationResourceProperty action (Python: create_integration_resource_property)
  - UpdateIntegrationResourceProperty action (Python: update_integration_resource_property)
  - GetIntegrationResourceProperty action (Python: get_integration_resource_property)
  - UntagResource action (Python: untag_resource)
  - ListTagsForResource action (Python: list_tags_for_resource)
  - Exceptions
  - ResourceNotFoundException structure
  - InternalServerException structure
  - IntegrationAlreadyExistsFault structure
  - IntegrationConflictOperationFault structure
  - IntegrationQuotaExceededFault structure
  - KMSKeyNotAccessibleFault structure
  - IntegrationNotFoundFault structure
  - TargetResourceNotFound structure
  - InvalidIntegrationStateFault structure
- Interactive sessions API
  - Data types
  - Session structure
  - SessionCommand structure
  - Statement structure
  - StatementOutput structure
  - StatementOutputData structure
  - ConnectionsList structure
  - Operations
  - CreateSession action (Python: create_session)
  - StopSession action (Python: stop_session)
  - DeleteSession action (Python: delete_session)
  - GetSession action (Python: get_session)

  - ListSessions action (Python: list_sessions)

  - RunStatement action (Python: run_statement)
  - CancelStatement action (Python: cancel_statement)
  - GetStatement action (Python: get_statement)
  - ListStatements action (Python: list_statements)
  - GetGlueIdentityCenterConfiguration action (Python: get_glue_identity_center_configuration)
  - UpdateGlueIdentityCenterConfiguration action (Python: update_glue_identity_center_configuration)
  - CreateGlueIdentityCenterConfiguration action (Python: create_glue_identity_center_configuration)
  - DeleteGlueIdentityCenterConfiguration action (Python: delete_glue_identity_center_configuration)
- Development endpoints API
  - Data types
  - DevEndpoint structure
  - DevEndpointCustomLibraries structure
  - Operations
  - CreateDevEndpoint action (Python: create_dev_endpoint)
  - UpdateDevEndpoint action (Python: update_dev_endpoint)
  - DeleteDevEndpoint action (Python: delete_dev_endpoint)
  - GetDevEndpoint action (Python: get_dev_endpoint)
  - GetDevEndpoints action (Python: get_dev_endpoints)
  - BatchGetDevEndpoints action (Python: batch_get_dev_endpoints)
  - ListDevEndpoints action (Python: list_dev_endpoints)
- Schema registry
  - Data types
  - RegistryId structure
  - RegistryListItem structure
  - MetadataInfo structure
  - OtherMetadataValueListItem structure
  - SchemaListItem structure
  - SchemaVersionListItem structure

  - MetadataKeyValuePair structure
  - SchemaVersionErrorItem structure
  - ErrorDetails structure
  - SchemaVersionNumber structure
  - SchemaId structure
  - Operations
  - CreateRegistry action (Python: create_registry)
  - CreateSchema action (Python: create_schema)
  - GetSchema action (Python: get_schema)
  - ListSchemaVersions action (Python: list_schema_versions)
  - GetSchemaVersion action (Python: get_schema_version)
  - GetSchemaVersionsDiff action (Python: get_schema_versions_diff)
  - ListRegistries action (Python: list_registries)
  - ListSchemas action (Python: list_schemas)
  - RegisterSchemaVersion action (Python: register_schema_version)
  - UpdateSchema action (Python: update_schema)
  - CheckSchemaVersionValidity action (Python: check_schema_version_validity)
  - UpdateRegistry action (Python: update_registry)
  - GetSchemaByDefinition action (Python: get_schema_by_definition)
  - GetRegistry action (Python: get_registry)
  - PutSchemaVersionMetadata action (Python: put_schema_version_metadata)
  - QuerySchemaVersionMetadata action (Python: query_schema_version_metadata)
  - RemoveSchemaVersionMetadata action (Python: remove_schema_version_metadata)
  - DeleteRegistry action (Python: delete_registry)
  - DeleteSchema action (Python: delete_schema)
  - DeleteSchemaVersions action (Python: delete_schema_versions)
- Workflows
  - Data types
  - JobNodeDetails structure
  - CrawlerNodeDetails structure

- TriggerNodeDetails structure
- Crawl structure
- Node structure
- Edge structure
- Workflow structure
- WorkflowGraph structure
- WorkflowRun structure
- WorkflowRunStatistics structure
- StartingEventBatchCondition structure
- Blueprint structure
- BlueprintDetails structure
- LastActiveDefinition structure
- BlueprintRun structure
- Operations
- CreateWorkflow action (Python: create_workflow)
- UpdateWorkflow action (Python: update_workflow)
- DeleteWorkflow action (Python: delete_workflow)
- GetWorkflow action (Python: get_workflow)
- ListWorkflows action (Python: list_workflows)
- BatchGetWorkflows action (Python: batch_get_workflows)
- GetWorkflowRun action (Python: get_workflow_run)
- GetWorkflowRuns action (Python: get_workflow_runs)
- GetWorkflowRunProperties action (Python: get_workflow_run_properties)
- PutWorkflowRunProperties action (Python: put_workflow_run_properties)
- CreateBlueprint action (Python: create_blueprint)
- UpdateBlueprint action (Python: update_blueprint)
- DeleteBlueprint action (Python: delete_blueprint)
- ListBlueprints action (Python: list_blueprints)
- BatchGetBlueprints action (Python: batch_get_blueprints)

- StartBlueprintRun action (Python: start_blueprint_run)

  - GetBlueprintRun action (Python: get_blueprint_run)
  - GetBlueprintRuns action (Python: get_blueprint_runs)
  - StartWorkflowRun action (Python: start_workflow_run)
  - StopWorkflowRun action (Python: stop_workflow_run)
  - ResumeWorkflowRun action (Python: resume_workflow_run)
- Usage profiles
  - Data types
  - ProfileConfiguration structure
  - ConfigurationObject structure
  - UsageProfileDefinition structure
  - Operations
  - CreateUsageProfile action (Python: create_usage_profile)
  - GetUsageProfile action (Python: get_usage_profile)
  - UpdateUsageProfile action (Python: update_usage_profile)
  - DeleteUsageProfile action (Python: delete_usage_profile)
  - ListUsageProfiles action (Python: list_usage_profiles)
- Machine learning API
  - Data types
  - TransformParameters structure
  - EvaluationMetrics structure
  - MLTransform structure
  - FindMatchesParameters structure
  - FindMatchesMetrics structure
  - ConfusionMatrix structure
  - GlueTable structure
  - TaskRun structure
  - TransformFilterCriteria structure
  - TransformSortCriteria structure
  - TaskRunFilterCriteria structure
  - TaskRunSortCriteria structure

  - TaskRunProperties structure
  - FindMatchesTaskRunProperties structure
  - ImportLabelsTaskRunProperties structure
  - ExportLabelsTaskRunProperties structure
  - LabelingSetGenerationTaskRunProperties structure
  - SchemaColumn structure
  - TransformEncryption structure
  - MLUserDataEncryption structure
  - ColumnImportance structure
  - Operations
  - CreateMLTransform action (Python: create_ml_transform)
  - UpdateMLTransform action (Python: update_ml_transform)
  - DeleteMLTransform action (Python: delete_ml_transform)
  - GetMLTransform action (Python: get_ml_transform)
  - GetMLTransforms action (Python: get_ml_transforms)
  - ListMLTransforms action (Python: list_ml_transforms)
  - StartMLEvaluationTaskRun action (Python: start_ml_evaluation_task_run)
  - StartMLLabelingSetGenerationTaskRun action (Python: start_ml_labeling_set_generation_task_run)
  - GetMLTaskRun action (Python: get_ml_task_run)
  - GetMLTaskRuns action (Python: get_ml_task_runs)
  - CancelMLTaskRun action (Python: cancel_ml_task_run)
  - StartExportLabelsTaskRun action (Python: start_export_labels_task_run)
  - StartImportLabelsTaskRun action (Python: start_import_labels_task_run)
- Data Quality API
  - Data types
  - DataSource structure
  - DataQualityRulesetListDetails structure
  - DataQualityTargetTable structure

  - DataQualityRulesetEvaluationRunDescription structure

- DataQualityRulesetEvaluationRunFilter structure
- DataQualityEvaluationRunAdditionalRunOptions structure
- DataQualityRuleRecommendationRunDescription structure
- DataQualityRuleRecommendationRunFilter structure
- DataQualityResult structure
- DataQualityAnalyzerResult structure
- DataQualityObservation structure
- MetricBasedObservation structure
- DataQualityMetricValues structure
- DataQualityRuleResult structure
- DataQualityResultDescription structure
- DataQualityResultFilterCriteria structure
- DataQualityRulesetFilterCriteria structure
- DataQualityAggregatedMetrics structure
- StatisticAnnotation structure
- TimestampedInclusionAnnotation structure
- AnnotationError structure
- DatapointInclusionAnnotation structure
- StatisticSummaryList list
- StatisticSummary structure
- RunIdentifier structure
- StatisticModelResult structure
- DataQualityGlueTable structure
- Operations
- StartDataQualityRulesetEvaluationRun action (Python: start_data_quality_ruleset_evaluation_run)
- CancelDataQualityRulesetEvaluationRun action (Python: cancel_data_quality_ruleset_evaluation_run)
- GetDataQualityRulesetEvaluationRun action (Python: get_data_quality_ruleset_evaluation_run)

- ListDataQualityRulesetEvaluationRuns action (Python: list_data_quality_ruleset_evaluation_runs)
- StartDataQualityRuleRecommendationRun action (Python: start_data_quality_rule_recommendation_run)
- CancelDataQualityRuleRecommendationRun action (Python: cancel_data_quality_rule_recommendation_run)
- GetDataQualityRuleRecommendationRun action (Python: get_data_quality_rule_recommendation_run)
- ListDataQualityRuleRecommendationRuns action (Python: list_data_quality_rule_recommendation_runs)
- GetDataQualityResult action (Python: get_data_quality_result)
- BatchGetDataQualityResult action (Python: batch_get_data_quality_result)
- ListDataQualityResults action (Python: list_data_quality_results)
- CreateDataQualityRuleset action (Python: create_data_quality_ruleset)
- DeleteDataQualityRuleset action (Python: delete_data_quality_ruleset)
- GetDataQualityRuleset action (Python: get_data_quality_ruleset)
- ListDataQualityRulesets action (Python: list_data_quality_rulesets)
- UpdateDataQualityRuleset action (Python: update_data_quality_ruleset)
- ListDataQualityStatistics action (Python: list_data_quality_statistics)
- TimestampFilter structure
- CreateDataQualityRulesetRequest structure
- GetDataQualityRulesetResponse structure
- GetDataQualityResultResponse structure
- StartDataQualityRuleRecommendationRunRequest structure
- GetDataQualityRuleRecommendationRunResponse structure
- BatchPutDataQualityStatisticAnnotation action (Python: batch_put_data_quality_statistic_annotation)
- GetDataQualityModel action (Python: get_data_quality_model)
- GetDataQualityModelResult action (Python: get_data_quality_model_result)
- ListDataQualityStatisticAnnotations action (Python: list_data_quality_statistic_annotations)
- PutDataQualityProfileAnnotation action (Python: put_data_quality_profile_annotation)

- Sensitive data detection API
  - Data types
  - CustomEntityType structure
  - Operations
  - CreateCustomEntityType action (Python: create_custom_entity_type)
  - DeleteCustomEntityType action (Python: delete_custom_entity_type)
  - GetCustomEntityType action (Python: get_custom_entity_type)
  - BatchGetCustomEntityTypes action (Python: batch_get_custom_entity_types)
  - ListCustomEntityTypes action (Python: list_custom_entity_types)
- Tagging APIs in AWS Glue
  - Data types
  - Tag structure
  - Operations
  - TagResource action (Python: tag_resource)
  - UntagResource action (Python: untag_resource)
  - GetTags action (Python: get_tags)
- Common data types
  - Tag structure
  - DecimalNumber structure
  - ErrorDetail structure
  - PropertyPredicate structure
  - ResourceUri structure
  - ColumnStatistics structure
  - ColumnStatisticsError structure
  - ColumnError structure
  - ColumnStatisticsData structure
  - BooleanColumnStatisticsData structure
  - DateColumnStatisticsData structure
  - DecimalColumnStatisticsData structure
  - DoubleColumnStatisticsData structure

  - LongColumnStatisticsData structure
  - StringColumnStatisticsData structure
  - BinaryColumnStatisticsData structure
  - String patterns
- Exceptions
  - AccessDeniedException structure
  - AlreadyExistsException structure
  - ConcurrentModificationException structure
  - ConcurrentRunsExceededException structure
  - CrawlerNotRunningException structure
  - CrawlerRunningException structure
  - CrawlerStoppingException structure
  - EntityNotFoundException structure
  - FederationSourceException structure
  - FederationSourceRetryableException structure
  - GlueEncryptionException structure
  - IdempotentParameterMismatchException structure
  - IllegalWorkflowStateException structure
  - InternalServiceException structure
  - InvalidExecutionEngineException structure
  - InvalidInputException structure
  - InvalidStateException structure
  - InvalidTaskStatusTransitionException structure
  - JobDefinitionErrorException structure
  - JobRunInTerminalStateException structure
  - JobRunInvalidStateTransitionException structure
  - JobRunNotInTerminalStateException structure
  - LateRunnerException structure
  - NoScheduleException structure

  - OperationTimeoutException structure

  - ResourceNotReadyException structure
  - ResourceNumberLimitExceededException structure
  - SchedulerNotRunningException structure
  - SchedulerRunningException structure
  - SchedulerTransitioningException structure
  - UnrecognizedRunnerException structure
  - ValidationException structure
  - VersionMismatchException structure

Security APIs in AWS Glue

The Security API describes the security data types, and the API related to security in AWS Glue.

Data types

- DataCatalogEncryptionSettings structure
- EncryptionAtRest structure
- ConnectionPasswordEncryption structure
- EncryptionConfiguration structure
- S3Encryption structure
- CloudWatchEncryption structure
- JobBookmarksEncryption structure
- SecurityConfiguration structure
- GluePolicy structure
- DataQualityEncryption structure

DataCatalogEncryptionSettings structure

Contains configuration information for maintaining Data Catalog security.

Fields

- `EncryptionAtRest` – An EncryptionAtRest object.

Specifies the encryption-at-rest configuration for the Data Catalog.

- `ConnectionPasswordEncryption` – A ConnectionPasswordEncryption object.

  When connection password protection is enabled, the Data Catalog uses a customer-provided key to encrypt the password as part of `CreateConnection` or `UpdateConnection` and store it in the `ENCRYPTED_PASSWORD` field in the connection properties. You can enable catalog encryption or only password encryption.

EncryptionAtRest structure

Specifies the encryption-at-rest configuration for the Data Catalog.

Fields

- `CatalogEncryptionMode` – *Required:* UTF-8 string (valid values: `DISABLED` | `SSE-KMS="SSEKMS"` | `SSE-KMS-WITH-SERVICE-ROLE="SSEKMSWITHSERVICEROLE"`).

  The encryption-at-rest mode for encrypting Data Catalog data.
- `SseAwsKmsKeyId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the AWS KMS key to use for encryption at rest.
- `CatalogEncryptionServiceRole` – UTF-8 string, matching the Custom string pattern #51.

  The role that AWS Glue assumes to encrypt and decrypt the Data Catalog objects on the caller's behalf.

ConnectionPasswordEncryption structure

The data structure used by the Data Catalog to encrypt the password as part of `CreateConnection` or `UpdateConnection` and store it in the `ENCRYPTED_PASSWORD` field in the connection properties. You can enable catalog encryption or only password encryption.

When a `CreationConnection` request arrives containing a password, the Data Catalog first encrypts the password using your AWS KMS key. It then encrypts the whole connection object again if catalog encryption is also enabled.

This encryption requires that you set AWS KMS key permissions to enable or restrict access on the password key according to your security requirements. For example, you might want only administrators to have decrypt permission on the password key.

Fields

- `ReturnConnectionPasswordEncrypted` – *Required:* Boolean.

  When the `ReturnConnectionPasswordEncrypted` flag is set to "true", passwords remain encrypted in the responses of `GetConnection` and `GetConnections`. This encryption takes effect independently from catalog encryption.
- `AwsKmsKeyId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  An AWS KMS key that is used to encrypt the connection password.

  If connection password protection is enabled, the caller of `CreateConnection` and `UpdateConnection` needs at least `kms:Encrypt` permission on the specified AWS KMS key, to encrypt passwords before storing them in the Data Catalog.

  You can set the decrypt permission to enable or restrict access on the password key according to your security requirements.

EncryptionConfiguration structure

Specifies an encryption configuration.

Fields

- `S3Encryption` – An array of S3Encryption objects.

  The encryption configuration for Amazon Simple Storage Service (Amazon S3) data.
- `CloudWatchEncryption` – A CloudWatchEncryption object.

  The encryption configuration for Amazon CloudWatch.
- `JobBookmarksEncryption` – A JobBookmarksEncryption object.

  The encryption configuration for job bookmarks.
- `DataQualityEncryption` – A DataQualityEncryption object.

The encryption configuration for AWS Glue Data Quality assets.

S3Encryption structure

Specifies how Amazon Simple Storage Service (Amazon S3) data should be encrypted.

Fields

- `S3EncryptionMode` – UTF-8 string (valid values: `DISABLED` | `SSE-KMS="SSEKMS"` | `SSE-S3="SSES3"`).

  The encryption mode to use for Amazon S3 data.
- `KmsKeyArn` – UTF-8 string, matching the Custom string pattern #29.

  The Amazon Resource Name (ARN) of the KMS key to be used to encrypt the data.

CloudWatchEncryption structure

Specifies how Amazon CloudWatch data should be encrypted.

Fields

- `CloudWatchEncryptionMode` – UTF-8 string (valid values: `DISABLED` | `SSE-KMS="SSEKMS"`).

  The encryption mode to use for CloudWatch data.
- `KmsKeyArn` – UTF-8 string, matching the Custom string pattern #29.

  The Amazon Resource Name (ARN) of the KMS key to be used to encrypt the data.

JobBookmarksEncryption structure

Specifies how job bookmark data should be encrypted.

Fields

- `JobBookmarksEncryptionMode` – UTF-8 string (valid values: `DISABLED` | `CSE-KMS="CSEKMS"`).

  The encryption mode to use for job bookmarks data.

- `KmsKeyArn` – UTF-8 string, matching the Custom string pattern #29.

  The Amazon Resource Name (ARN) of the KMS key to be used to encrypt the data.

SecurityConfiguration structure

Specifies a security configuration.

Fields

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the security configuration.
- `CreatedTimeStamp` – Timestamp.

  The time at which this security configuration was created.
- `EncryptionConfiguration` – An EncryptionConfiguration object.

  The encryption configuration associated with this security configuration.

GluePolicy structure

A structure for returning a resource policy.

Fields

- `PolicyInJson` – UTF-8 string, at least 2 bytes long.

  Contains the requested policy document, in JSON format.
- `PolicyHash` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Contains the hash value associated with this policy.
- `CreateTime` – Timestamp.

  The date and time at which the policy was created.
- `UpdateTime` – Timestamp.

The date and time at which the policy was last updated.

DataQualityEncryption structure

Specifies how Data Quality assets in your account should be encrypted.

Fields

- `DataQualityEncryptionMode` – UTF-8 string (valid values: `DISABLED` | `SSE-KMS="SSEKMS"`).

  The encryption mode to use for encrypting Data Quality assets. These assets include data quality rulesets, results, statistics, anomaly detection models and observations.

  Valid values are `SSEKMS` for encryption using a customer-managed KMS key, or `DISABLED`.
- `KmsKeyArn` – UTF-8 string, matching the Custom string pattern #29.

  The Amazon Resource Name (ARN) of the KMS key to be used to encrypt the data.

Operations

- GetDataCatalogEncryptionSettings action (Python: get_data_catalog_encryption_settings)
- PutDataCatalogEncryptionSettings action (Python: put_data_catalog_encryption_settings)
- PutResourcePolicy action (Python: put_resource_policy)
- GetResourcePolicy action (Python: get_resource_policy)
- DeleteResourcePolicy action (Python: delete_resource_policy)
- CreateSecurityConfiguration action (Python: create_security_configuration)
- DeleteSecurityConfiguration action (Python: delete_security_configuration)
- GetSecurityConfiguration action (Python: get_security_configuration)
- GetSecurityConfigurations action (Python: get_security_configurations)
- GetResourcePolicies action (Python: get_resource_policies)

GetDataCatalogEncryptionSettings action (Python: get_data_catalog_encryption_settings)

Retrieves the security configuration for a specified catalog.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog to retrieve the security configuration for. If none is provided, the AWS account ID is used by default.

Response

- `DataCatalogEncryptionSettings` – A DataCatalogEncryptionSettings object.

  The requested security configuration.

Errors

- `InternalServiceException`
- `InvalidInputException`
- `OperationTimeoutException`

PutDataCatalogEncryptionSettings action (Python: put_data_catalog_encryption_settings)

Sets the security configuration for a specified catalog. After the configuration has been set, the specified encryption is applied to every catalog write thereafter.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog to set the security configuration for. If none is provided, the AWS account ID is used by default.

- `DataCatalogEncryptionSettings` – *Required:* A DataCatalogEncryptionSettings object.

  The security configuration to set.

Response

- *No Response parameters.*

Errors

- `InternalServiceException`
- `InvalidInputException`
- `OperationTimeoutException`

PutResourcePolicy action (Python: put_resource_policy)

Sets the Data Catalog resource policy for access control.

Request

- `PolicyInJson` – *Required:* UTF-8 string, at least 2 bytes long.

  Contains the policy document to set, in JSON format.
- `ResourceArn` – UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

  Do not use. For internal use only.
- `PolicyHashCondition` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The hash value returned when the previous policy was set using `PutResourcePolicy`. Its purpose is to prevent concurrent modifications of a policy. Do not use this parameter if no previous policy has been set.
- `PolicyExistsCondition` – UTF-8 string (valid values: `MUST_EXIST` | `NOT_EXIST` | `NONE`).

  A value of `MUST_EXIST` is used to update a policy. A value of `NOT_EXIST` is used to create a new policy. If a value of `NONE` or a null value is used, the call does not depend on the existence of a policy.

- `EnableHybrid` – UTF-8 string (valid values: `TRUE` | `FALSE`).

  If `'TRUE'`, indicates that you are using both methods to grant cross-account access to Data Catalog resources:

  - By directly updating the resource policy with `PutResourePolicy`
  - By using the **Grant permissions** command on the AWS Management Console.

  Must be set to `'TRUE'` if you have already used the Management Console to grant cross-account access, otherwise the call fails. Default is 'FALSE'.

Response

- `PolicyHash` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A hash of the policy that has just been set. This must be included in a subsequent call that overwrites or updates this policy.

Errors

- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`
- `InvalidInputException`
- `ConditionCheckFailureException`

GetResourcePolicy action (Python: get_resource_policy)

Retrieves a specified resource policy.

Request

- `ResourceArn` – UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

The ARN of the AWS Glue resource for which to retrieve the resource policy. If not supplied, the Data Catalog resource policy is returned. Use `GetResourcePolicies` to view all existing resource policies. For more information see Specifying AWS Glue Resource ARNs.

Response

- `PolicyInJson` – UTF-8 string, at least 2 bytes long.

  Contains the requested policy document, in JSON format.
- `PolicyHash` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Contains the hash value associated with this policy.
- `CreateTime` – Timestamp.

  The date and time at which the policy was created.
- `UpdateTime` – Timestamp.

  The date and time at which the policy was last updated.

Errors

- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`
- `InvalidInputException`

DeleteResourcePolicy action (Python: delete_resource_policy)

Deletes a specified policy.

Request

- `PolicyHashCondition` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The hash value returned when this policy was set.

- `ResourceArn` – UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

  The ARN of the AWS Glue resource for the resource policy to be deleted.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`
- `InvalidInputException`
- `ConditionCheckFailureException`

CreateSecurityConfiguration action (Python: create_security_configuration)

Creates a new security configuration. A security configuration is a set of security properties that can be used by AWS Glue. You can use a security configuration to encrypt data at rest. For information about using security configurations in AWS Glue, see Encrypting Data Written by Crawlers, Jobs, and Development Endpoints.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name for the new security configuration.
- `EncryptionConfiguration` – *Required:* An EncryptionConfiguration object.

  The encryption configuration for the new security configuration.

Response

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name assigned to the new security configuration.
- `CreatedTimestamp` – Timestamp.

  The time at which the new security configuration was created.

Errors

- `AlreadyExistsException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `ResourceNumberLimitExceededException`

DeleteSecurityConfiguration action (Python: delete_security_configuration)

Deletes a specified security configuration.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the security configuration to delete.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`

- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`

GetSecurityConfiguration action (Python: get_security_configuration)

Retrieves a specified security configuration.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the security configuration to retrieve.

Response

- `SecurityConfiguration` – A SecurityConfiguration object.

  The requested security configuration.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`

GetSecurityConfigurations action (Python: get_security_configurations)

Retrieves a list of all security configurations.

Request

- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum number of results to return.

- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation call.

Response

- `SecurityConfigurations` – An array of SecurityConfiguration objects.

  A list of security configurations.
- `NextToken` – UTF-8 string.

  A continuation token, if there are more security configurations to return.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`

GetResourcePolicies action (Python: get_resource_policies)

Retrieves the resource policies set on individual resources by AWS Resource Access Manager during cross-account permission grants. Also retrieves the Data Catalog resource policy.

If you enabled metadata encryption in Data Catalog settings, and you do not have permission on the AWS KMS key, the operation can't return the Data Catalog resource policy.

Request

- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation request.
- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum size of a list to return.

Response

- `GetResourcePoliciesResponseList` – An array of GluePolicy objects.

  A list of the individual resource policies and the account-level resource policy.
- `NextToken` – UTF-8 string.

  A continuation token, if the returned list does not contain the last resource policy available.

Errors

- `InternalServiceException`
- `OperationTimeoutException`
- `InvalidInputException`
- `GlueEncryptionException`

Catalog objects API

The Catalog objects API describes the data types and API related to working with catalogs in AWS Glue.

Topics

- Catalogs API
- Database API
- Table API
- Partition API
- Connections API
- User-defined Function API
- Importing an Athena catalog to AWS Glue

Catalogs API

The Catalogs API describes the APIs for creating, deleting, locating, updating, and listing catalogs.

Data types

- Catalog structure
- CatalogInput structure
- TargetRedshiftCatalog structure
- CatalogProperties structure
- CatalogPropertiesOutput structure
- DataLakeAccessProperties structure
- IcebergOptimizationProperties structure
- DataLakeAccessPropertiesOutput structure
- IcebergOptimizationPropertiesOutput structure
- FederatedCatalog structure

Catalog structure

The catalog object represents a logical grouping of databases in the AWS Glue Data Catalog or a federated source. You can now create a Redshift-federated catalog or a catalog containing resource links to Redshift databases in another account or region.

Fields

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the catalog. To grant access to the default catalog, this field should not be provided.

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 64 bytes long, matching the Custom string pattern #25.

  The name of the catalog. Cannot be the same as the account ID.

- `ResourceArn` – UTF-8 string.

  The Amazon Resource Name (ARN) assigned to the catalog resource.

- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern. A description of the catalog.

- `Parameters` – A map array of key-value pairs.

  Each key is a Key string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a UTF-8 string, not more than 512000 bytes long.

  A map array of key-value pairs that define parameters and properties of the catalog.
- `CreateTime` – Timestamp.

  The time at which the catalog was created.
- `UpdateTime` – Timestamp.

  The time at which the catalog was last updated.
- `TargetRedshiftCatalog` – A TargetRedshiftCatalog object.

  A `TargetRedshiftCatalog` object that describes a target catalog for database resource linking.
- `FederatedCatalog` – A FederatedCatalog object.

  A `FederatedCatalog` object that points to an entity outside the AWS Glue Data Catalog.
- `CatalogProperties` – A CatalogPropertiesOutput object.

  A `CatalogProperties` object that specifies data lake access properties and other custom properties.
- `CreateTableDefaultPermissions` – An array of PrincipalPermissions objects.

  An array of `PrincipalPermissions` objects. Creates a set of default permissions on the table(s) for principals. Used by AWS Lake Formation. Not used in the normal course of AWS Glue operations.
- `CreateDatabaseDefaultPermissions` – An array of PrincipalPermissions objects.

  An array of `PrincipalPermissions` objects. Creates a set of default permissions on the database(s) for principals. Used by AWS Lake Formation. Not used in the normal course of AWS Glue operations.

- `AllowFullTableExternalDataAccess` – UTF-8 string (valid values: `True` | `False`).

  Allows third-party engines to access data in Amazon S3 locations that are registered with Lake Formation.

CatalogInput structure

A structure that describes catalog properties.

Fields

- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern. A description of the catalog.
- `FederatedCatalog` – A FederatedCatalog object.

  A `FederatedCatalog` object. A `FederatedCatalog` structure that references an entity outside the AWS Glue Data Catalog, for example a Redshift database.
- `Parameters` – A map array of key-value pairs.

  Each key is a Key string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a UTF-8 string, not more than 512000 bytes long.

  A map array of key-value pairs that define the parameters and properties of the catalog.
- `TargetRedshiftCatalog` – A TargetRedshiftCatalog object.

  A `TargetRedshiftCatalog` object that describes a target catalog for resource linking.
- `CatalogProperties` – A CatalogProperties object.

  A `CatalogProperties` object that specifies data lake access properties and other custom properties.
- `CreateTableDefaultPermissions` – An array of PrincipalPermissions objects.

An array of `PrincipalPermissions` objects. Creates a set of default permissions on the table(s) for principals. Used by AWS Lake Formation. Typically should be explicitly set as an empty list.

- `CreateDatabaseDefaultPermissions` – An array of PrincipalPermissions objects.

  An array of `PrincipalPermissions` objects. Creates a set of default permissions on the database(s) for principals. Used by AWS Lake Formation. Typically should be explicitly set as an empty list.
- `AllowFullTableExternalDataAccess` – UTF-8 string (valid values: `True` | `False`).

  Allows third-party engines to access data in Amazon S3 locations that are registered with Lake Formation.

TargetRedshiftCatalog structure

A structure that describes a target catalog for resource linking.

Fields

- `CatalogArn` – *Required:* UTF-8 string.

  The Amazon Resource Name (ARN) of the catalog resource.

CatalogProperties structure

A structure that specifies data lake access properties and other custom properties.

Fields

- `DataLakeAccessProperties` – A DataLakeAccessProperties object.

  A `DataLakeAccessProperties` object that specifies properties to configure data lake access for your catalog resource in the AWS Glue Data Catalog.
- `IcebergOptimizationProperties` – An IcebergOptimizationProperties object.

  A structure that specifies Iceberg table optimization properties for the catalog. This includes configuration for compaction, retention, and orphan file deletion operations that can be applied to Iceberg tables in this catalog.

- `CustomProperties` – A map array of key-value pairs.

  Each key is a Key string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a UTF-8 string, not more than 512000 bytes long.

  Additional key-value properties for the catalog, such as column statistics optimizations.

CatalogPropertiesOutput structure

Property attributes that include configuration properties for the catalog resource.

Fields

- `DataLakeAccessProperties` – A DataLakeAccessPropertiesOutput object.

  A `DataLakeAccessProperties` object with input properties to configure data lake access for your catalog resource in the AWS Glue Data Catalog.
- `IcebergOptimizationProperties` – An IcebergOptimizationPropertiesOutput object.

  An `IcebergOptimizationPropertiesOutput` object that specifies Iceberg table optimization settings for the catalog, including configurations for compaction, retention, and orphan file deletion operations.
- `CustomProperties` – A map array of key-value pairs.

  Each key is a Key string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a UTF-8 string, not more than 512000 bytes long.

  Additional key-value properties for the catalog, such as column statistics optimizations.

DataLakeAccessProperties structure

Input properties to configure data lake access for your catalog resource in the AWS Glue Data Catalog.

Fields

- `DataLakeAccess` – Boolean.

  Turns on or off data lake access for Apache Spark applications that access Amazon Redshift databases in the Data Catalog from any non-Redshift engine, such as Amazon Athena, Amazon EMR, or AWS Glue ETL.
- `DataTransferRole` – UTF-8 string, matching the Custom string pattern #51.

  A role that will be assumed by AWS Glue for transferring data into/out of the staging bucket during a query.
- `KmsKey` – UTF-8 string.

  An encryption key that will be used for the staging bucket that will be created along with the catalog.
- `CatalogType` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Specifies a federated catalog type for the native catalog resource. The currently supported type is `aws:redshift`.

IcebergOptimizationProperties structure

A structure that specifies Iceberg table optimization properties for the catalog, including configurations for compaction, retention, and orphan file deletion operations.

Fields

- `RoleArn` – UTF-8 string, matching the Custom string pattern #51.

  The Amazon Resource Name (ARN) of the IAM role that will be assumed to perform Iceberg table optimization operations.
- `Compaction` – A map array of key-value pairs.

  Each key is a Key string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a UTF-8 string, not more than 512000 bytes long.

A map of key-value pairs that specify configuration parameters for Iceberg table compaction operations, which optimize the layout of data files to improve query performance.

- `Retention` – A map array of key-value pairs.

  Each key is a Key string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a UTF-8 string, not more than 512000 bytes long.

  A map of key-value pairs that specify configuration parameters for Iceberg table retention operations, which manage the lifecycle of table snapshots to control storage costs.
- `OrphanFileDeletion` – A map array of key-value pairs.

  Each key is a Key string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a UTF-8 string, not more than 512000 bytes long.

  A map of key-value pairs that specify configuration parameters for Iceberg orphan file deletion operations, which identify and remove files that are no longer referenced by the table metadata.

DataLakeAccessPropertiesOutput structure

The output properties of the data lake access configuration for your catalog resource in the AWS Glue Data Catalog.

Fields

- `DataLakeAccess` – Boolean.

  Turns on or off data lake access for Apache Spark applications that access Amazon Redshift databases in the Data Catalog.
- `DataTransferRole` – UTF-8 string, matching the Custom string pattern #51.

  A role that will be assumed by AWS Glue for transferring data into/out of the staging bucket during a query.
- `KmsKey` – UTF-8 string.

An encryption key that will be used for the staging bucket that will be created along with the catalog.

- `ManagedWorkgroupName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The managed Redshift Serverless compute name that is created for your catalog resource.
- `ManagedWorkgroupStatus` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The managed Redshift Serverless compute status.
- `RedshiftDatabaseName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The default Redshift database resource name in the managed compute.
- `StatusMessage` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A message that gives more detailed information about the managed workgroup status.
- `CatalogType` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Specifies a federated catalog type for the native catalog resource. The currently supported type is `aws:redshift`.

IcebergOptimizationPropertiesOutput structure

A structure that contains the output properties of Iceberg table optimization configuration for your catalog resource in the AWS Glue Data Catalog.

Fields

- `RoleArn` – UTF-8 string, matching the Custom string pattern #51.

  The Amazon Resource Name (ARN) of the IAM role that is used to perform Iceberg table optimization operations.
- `Compaction` – A map array of key-value pairs.

Each key is a Key string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

Each value is a UTF-8 string, not more than 512000 bytes long.

A map of key-value pairs that specify configuration parameters for Iceberg table compaction operations, which optimize the layout of data files to improve query performance.

- `Retention` – A map array of key-value pairs.

  Each key is a Key string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a UTF-8 string, not more than 512000 bytes long.

  A map of key-value pairs that specify configuration parameters for Iceberg table retention operations, which manage the lifecycle of table snapshots to control storage costs.
- `OrphanFileDeletion` – A map array of key-value pairs.

  Each key is a Key string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a UTF-8 string, not more than 512000 bytes long.

  A map of key-value pairs that specify configuration parameters for Iceberg orphan file deletion operations, which identify and remove files that are no longer referenced by the table metadata.
- `LastUpdatedTime` – Timestamp.

  The timestamp when the Iceberg optimization properties were last updated.

FederatedCatalog structure

A catalog that points to an entity outside the AWS Glue Data Catalog.

Fields

- `Identifier` – UTF-8 string, not less than 1 or more than 512 bytes long, matching the Single-line string pattern.

  A unique identifier for the federated catalog.

- `ConnectionName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the connection to an external data source, for example a Redshift-federated catalog.

- `ConnectionType` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The type of connection used to access the federated catalog, specifying the protocol or method for connection to the external data source.

Operations

- CreateCatalog action (Python: create_catalog)
- UpdateCatalog action (Python: update_catalog)
- DeleteCatalog action (Python: delete_catalog)
- GetCatalog action (Python: get_catalog)
- GetCatalogs action (Python: get_catalogs)

CreateCatalog action (Python: create_catalog)

Creates a new catalog in the AWS Glue Data Catalog.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 64 bytes long, matching the Custom string pattern #25.

  The name of the catalog to create.

- `CatalogInput` – *Required:* A CatalogInput object.

  A `CatalogInput` object that defines the metadata for the catalog.

- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

A map array of key-value pairs, not more than 50 pairs. Each key is a UTF-8 string, not less than 1 or more than 128 bytes long. Each value is a UTF-8 string, not more than 256 bytes long. The tags you assign to the catalog.

Response

- *No Response parameters.*

Errors

- `InvalidInputException`
- `AlreadyExistsException`
- `ResourceNumberLimitExceededException`
- `InternalServiceException`
- `OperationTimeoutException`
- `GlueEncryptionException`
- `ConcurrentModificationException`
- `AccessDeniedException`
- `EntityNotFoundException`
- `FederatedResourceAlreadyExistsException`
- `FederationSourceException`

UpdateCatalog action (Python: update_catalog)

Updates an existing catalog's properties in the AWS Glue Data Catalog.

Request

- `CatalogId` – *Required:* Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the catalog.
- `CatalogInput` – *Required:* A CatalogInput object.

  A `CatalogInput` object specifying the new properties of an existing catalog.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `GlueEncryptionException`
- `ConcurrentModificationException`
- `AccessDeniedException`
- `FederationSourceException`

DeleteCatalog action (Python: delete_catalog)

Removes the specified catalog from the AWS Glue Data Catalog.

After completing this operation, you no longer have access to the databases, tables (and all table versions and partitions that might belong to the tables) and the user-defined functions in the deleted catalog. AWS Glue deletes these "orphaned" resources asynchronously in a timely manner, at the discretion of the service.

To ensure the immediate deletion of all related resources before calling the `DeleteCatalog` operation, use `DeleteTableVersion` (or `BatchDeleteTableVersion`), `DeletePartition` (or `BatchDeletePartition`), `DeleteTable` (or `BatchDeleteTable`), `DeleteUserDefinedFunction` and `DeleteDatabase` to delete any resources that belong to the catalog.

Request

- `CatalogId` – *Required:* Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the catalog.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `GlueEncryptionException`
- `ConcurrentModificationException`
- `AccessDeniedException`
- `FederationSourceException`

GetCatalog action (Python: get_catalog)

The name of the Catalog to retrieve. This should be all lowercase.

Request

- `CatalogId` – *Required:* Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the parent catalog in which the catalog resides. If none is provided, the AWS Account Number is used by default.

Response

- `Catalog` – A Catalog object.

  A `Catalog` object. The definition of the specified catalog in the AWS Glue Data Catalog.

Errors

- `InvalidInputException`

- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`
- `GlueEncryptionException`
- `AccessDeniedException`
- `EntityNotFoundException`
- `FederationSourceException`
- `FederationSourceRetryableException`

GetCatalogs action (Python: get_catalogs)

Retrieves all catalogs defined in a catalog in the AWS Glue Data Catalog. For a Redshift-federated catalog use case, this operation returns the list of catalogs mapped to Redshift databases in the Redshift namespace catalog.

Request

- `ParentCatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the parent catalog in which the catalog resides. If none is provided, the AWS Account Number is used by default.

- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation call.

- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum number of catalogs to return in one response.

- `Recursive` – Boolean.

  Whether to list all catalogs across the catalog hierarchy, starting from the `ParentCatalogId`. Defaults to `false` . When `true`, all catalog objects in the `ParentCatalogID` hierarchy are enumerated in the response.

- `IncludeRoot` – Boolean.

Whether to list the default catalog in the account and region in the response. Defaults to `false`. When `true` and `ParentCatalogId = NULL | AWS Account ID`, all catalogs and the default catalog are enumerated in the response.

When the `ParentCatalogId` is not equal to null, and this attribute is passed as `false` or `true`, an `InvalidInputException` is thrown.

Response

- `CatalogList` – *Required:* An array of Catalog objects.

  An array of `Catalog` objects. A list of `Catalog` objects from the specified parent catalog.
- `NextToken` – UTF-8 string.

  A continuation token for paginating the returned list of tokens, returned if the current segment of the list is not the last.

Errors

- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `GlueEncryptionException`
- `AccessDeniedException`
- `EntityNotFoundException`
- `FederationSourceException`
- `FederationSourceRetryableException`

Database API

The Database API describes database data types, and includes the API for creating, deleting, locating, updating, and listing databases.

Data types

- Database structure
- DatabaseInput structure
- PrincipalPermissions structure
- DataLakePrincipal structure
- DatabaseIdentifier structure
- FederatedDatabase structure

Database structure

The `Database` object represents a logical grouping of tables that might reside in a Hive metastore or an RDBMS.

Fields

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the database. For Hive compatibility, this is folded to lowercase when it is stored.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the database.
- `LocationUri` – Uniform resource identifier (uri), not less than 1 or more than 1024 bytes long, matching the URI address multi-line string pattern.

  The location of the database (for example, an HDFS path).
- `Parameters` – A map array of key-value pairs.

  Each key is a Key string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a UTF-8 string, not more than 512000 bytes long.

  These key-value pairs define parameters and properties of the database.
- `CreateTime` – Timestamp.

The time at which the metadata database was created in the catalog.

- `CreateTableDefaultPermissions` – An array of PrincipalPermissions objects.

  Creates a set of default permissions on the table for principals. Used by AWS Lake Formation. Not used in the normal course of AWS Glue operations.
- `TargetDatabase` – A DatabaseIdentifier object.

  A `DatabaseIdentifier` structure that describes a target database for resource linking.
- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog in which the database resides.
- `FederatedDatabase` – A FederatedDatabase object.

  A `FederatedDatabase` structure that references an entity outside the AWS Glue Data Catalog.

DatabaseInput structure

The structure used to create or update a database.

Fields

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the database. For Hive compatibility, this is folded to lowercase when it is stored.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the database.
- `LocationUri` – Uniform resource identifier (uri), not less than 1 or more than 1024 bytes long, matching the URI address multi-line string pattern.

  The location of the database (for example, an HDFS path).
- `Parameters` – A map array of key-value pairs.

  Each key is a Key string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

Each value is a UTF-8 string, not more than 512000 bytes long.

These key-value pairs define parameters and properties of the database.

These key-value pairs define parameters and properties of the database.

- `CreateTableDefaultPermissions` – An array of PrincipalPermissions objects.

  Creates a set of default permissions on the table for principals. Used by AWS Lake Formation. Not used in the normal course of AWS Glue operations.
- `TargetDatabase` – A DatabaseIdentifier object.

  A `DatabaseIdentifier` structure that describes a target database for resource linking.
- `FederatedDatabase` – A FederatedDatabase object.

  A `FederatedDatabase` structure that references an entity outside the AWS Glue Data Catalog.

PrincipalPermissions structure

Permissions granted to a principal.

Fields

- `Principal` – A DataLakePrincipal object.

  The principal who is granted permissions.
- `Permissions` – An array of UTF-8 strings.

  The permissions that are granted to the principal.

DataLakePrincipal structure

The AWS Lake Formation principal.

Fields

- `DataLakePrincipalIdentifier` – UTF-8 string, not less than 1 or more than 255 bytes long.

  An identifier for the AWS Lake Formation principal.

DatabaseIdentifier structure

A structure that describes a target database for resource linking.

Fields

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog in which the database resides.
- `DatabaseName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the catalog database.
- `Region` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Region of the target database.

FederatedDatabase structure

A database that points to an entity outside the AWS Glue Data Catalog.

Fields

- `Identifier` – UTF-8 string, not less than 1 or more than 512 bytes long, matching the Single-line string pattern.

  A unique identifier for the federated database.
- `ConnectionName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the connection to the external metastore.
- `ConnectionType` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The type of connection used to access the federated database, such as JDBC, ODBC, or other supported connection protocols.

Operations

- CreateDatabase action (Python: create_database)
- UpdateDatabase action (Python: update_database)
- DeleteDatabase action (Python: delete_database)
- GetDatabase action (Python: get_database)
- GetDatabases action (Python: get_databases)

CreateDatabase action (Python: create_database)

Creates a new database in a Data Catalog.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog in which to create the database. If none is provided, the AWS account ID is used by default.
- `DatabaseInput` – *Required:* A DatabaseInput object.

  The metadata for the database.
- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  The tags you assign to the database.

Response

- *No Response parameters.*

Errors

- `InvalidInputException`
- `AlreadyExistsException`

- `ResourceNumberLimitExceededException`
- `InternalServiceException`
- `OperationTimeoutException`
- `GlueEncryptionException`
- `ConcurrentModificationException`
- `FederatedResourceAlreadyExistsException`
- `FederationSourceException`
- `FederationSourceRetryableException`

UpdateDatabase action (Python: update_database)

Updates an existing database definition in a Data Catalog.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog in which the metadata database resides. If none is provided, the AWS account ID is used by default.
- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the database to update in the catalog. For Hive compatibility, this is folded to lowercase.
- `DatabaseInput` – *Required:* A DatabaseInput object.

  A `DatabaseInput` object specifying the new definition of the metadata database in the catalog.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`

- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `GlueEncryptionException`
- `ConcurrentModificationException`
- `FederationSourceException`
- `FederationSourceRetryableException`
- `AlreadyExistsException`

DeleteDatabase action (Python: delete_database)

Removes a specified database from a Data Catalog.

> **Note**
>
> After completing this operation, you no longer have access to the tables (and all table versions and partitions that might belong to the tables) and the user-defined functions in the deleted database. AWS Glue deletes these "orphaned" resources asynchronously in a timely manner, at the discretion of the service.
> To ensure the immediate deletion of all related resources, before calling `DeleteDatabase`, use `DeleteTableVersion` or `BatchDeleteTableVersion`, `DeletePartition` or `BatchDeletePartition`, `DeleteUserDefinedFunction`, and `DeleteTable` or `BatchDeleteTable`, to delete any resources that belong to the database.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog in which the database resides. If none is provided, the AWS account ID is used by default.

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the database to delete. For Hive compatibility, this must be all lowercase.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `ConcurrentModificationException`
- `FederationSourceException`
- `FederationSourceRetryableException`

GetDatabase action (Python: get_database)

Retrieves the definition of a specified database.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog in which the database resides. If none is provided, the AWS account ID is used by default.

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the database to retrieve. For Hive compatibility, this should be all lowercase.

Response

- `Database` – A Database object.

  The definition of the specified database in the Data Catalog.

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`
- `GlueEncryptionException`
- `FederationSourceException`
- `FederationSourceRetryableException`

GetDatabases action (Python: get_databases)

Retrieves all databases defined in a given Data Catalog.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog from which to retrieve `Databases`. If none is provided, the AWS account ID is used by default.
- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation call.
- `MaxResults` – Number (integer), not less than 1 or more than 100.

  The maximum number of databases to return in one response.
- `ResourceShareType` – UTF-8 string (valid values: `FOREIGN` | `ALL` | `FEDERATED`).

  Allows you to specify that you want to list the databases shared with your account. The allowable values are `FEDERATED`, `FOREIGN` or `ALL`.
  - If set to `FEDERATED`, will list the federated databases (referencing an external entity) shared with your account.
  - If set to `FOREIGN`, will list the databases shared with your account.
  - If set to `ALL`, will list the databases shared with your account, as well as the databases in yor local account.

- `AttributesToGet` – An array of UTF-8 strings.

  Specifies the database fields returned by the `GetDatabases` call. This parameter doesn't accept an empty list. The request must include the `NAME`.

Response

- `DatabaseList` – *Required:* An array of Database objects.

  A list of `Database` objects from the specified catalog.
- `NextToken` – UTF-8 string.

  A continuation token for paginating the returned list of tokens, returned if the current segment of the list is not the last.

Errors

- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `GlueEncryptionException`
- `EntityNotFoundException`
- `FederationSourceException`
- `FederationSourceRetryableException`

Table API

The Table API describes data types and operations associated with tables.

Data types

- Table structure
- TableInput structure
- FederatedTable structure
- Column structure
- StorageDescriptor structure

- SchemaReference structure
- SerDeInfo structure
- Order structure
- SkewedInfo structure
- TableVersion structure
- TableError structure
- TableVersionError structure
- SortCriterion structure
- TableIdentifier structure
- KeySchemaElement structure
- PartitionIndex structure
- PartitionIndexDescriptor structure
- BackfillError structure
- IcebergInput structure
- OpenTableFormatInput structure
- ViewDefinition structure
- ViewDefinitionInput structure
- ViewRepresentation structure
- ViewRepresentationInput structure
- UpdateOpenTableFormatInput structure
- UpdateIcebergInput structure
- CreateIcebergTableInput structure
- UpdateIcebergTableInput structure
- IcebergSortOrder structure
- IcebergSortField structure
- IcebergPartitionSpec structure
- IcebergPartitionField structure
- IcebergSchema structure
- IcebergStructField structure
- IcebergTableUpdate structure

Table structure

Represents a collection of related data organized in columns and rows.

Fields

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The table name. For Hive compatibility, this must be entirely lowercase.
- `DatabaseName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the database where the table metadata resides. For Hive compatibility, this must be all lowercase.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the table.
- `Owner` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The owner of the table.
- `CreateTime` – Timestamp.

  The time when the table definition was created in the Data Catalog.
- `UpdateTime` – Timestamp.

  The last time that the table was updated.
- `LastAccessTime` – Timestamp.

  The last time that the table was accessed. This is usually taken from HDFS, and might not be reliable.
- `LastAnalyzedTime` – Timestamp.

  The last time that column statistics were computed for this table.
- `Retention` – Number (integer), not more than None.

  The retention time for this table.

- `StorageDescriptor` – A StorageDescriptor object.

  A storage descriptor containing information about the physical storage of this table.

- `PartitionKeys` – An array of Column objects.

  A list of columns by which the table is partitioned. Only primitive types are supported as partition keys.

  When you create a table used by Amazon Athena, and you do not specify any `partitionKeys`, you must at least set the value of `partitionKeys` to an empty list. For example:

  ```
  "PartitionKeys": []
  ```

- `ViewOriginalText` – UTF-8 string, not more than 409600 bytes long.

  Included for Apache Hive compatibility. Not used in the normal course of AWS Glue operations. If the table is a `VIRTUAL_VIEW`, certain Athena configuration encoded in base64.

- `ViewExpandedText` – UTF-8 string, not more than 409600 bytes long.

  Included for Apache Hive compatibility. Not used in the normal course of AWS Glue operations.

- `TableType` – UTF-8 string, not more than 255 bytes long.

  The type of this table. AWS Glue will create tables with the `EXTERNAL_TABLE` type. Other services, such as Athena, may create tables with additional table types.

  AWS Glue related table types:

  EXTERNAL_TABLE

    Hive compatible attribute - indicates a non-Hive managed table.

  GOVERNED

    Used by AWS Lake Formation. The AWS Glue Data Catalog understands `GOVERNED`.

- `Parameters` – A map array of key-value pairs.

  Each key is a Key string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a UTF-8 string, not more than 512000 bytes long.

  These key-value pairs define properties associated with the table.

- `CreatedBy` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The person or entity who created the table.
- `IsRegisteredWithLakeFormation` – Boolean.

  Indicates whether the table has been registered with AWS Lake Formation.
- `TargetTable` – A TableIdentifier object.

  A `TableIdentifier` structure that describes a target table for resource linking.
- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog in which the table resides.
- `VersionId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the table version.
- `FederatedTable` – A FederatedTable object.

  A `FederatedTable` structure that references an entity outside the AWS Glue Data Catalog.
- `ViewDefinition` – A ViewDefinition object.

  A structure that contains all the information that defines the view, including the dialect or dialects for the view, and the query.
- `IsMultiDialectView` – Boolean.

  Specifies whether the view supports the SQL dialects of one or more different query engines and can therefore be read by those engines.

TableInput structure

A structure used to define a table.

Fields

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The table name. For Hive compatibility, this is folded to lowercase when it is stored.

- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the table.

- `Owner` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The table owner. Included for Apache Hive compatibility. Not used in the normal course of AWS Glue operations.

- `LastAccessTime` – Timestamp.

  The last time that the table was accessed.

- `LastAnalyzedTime` – Timestamp.

  The last time that column statistics were computed for this table.

- `Retention` – Number (integer), not more than None.

  The retention time for this table.

- `StorageDescriptor` – A StorageDescriptor object.

  A storage descriptor containing information about the physical storage of this table.

- `PartitionKeys` – An array of Column objects.

  A list of columns by which the table is partitioned. Only primitive types are supported as partition keys.

  When you create a table used by Amazon Athena, and you do not specify any `partitionKeys`, you must at least set the value of `partitionKeys` to an empty list. For example:

  ```
  "PartitionKeys": []
  ```

- `ViewOriginalText` – UTF-8 string, not more than 409600 bytes long.

  Included for Apache Hive compatibility. Not used in the normal course of AWS Glue operations. If the table is a `VIRTUAL_VIEW`, certain Athena configuration encoded in base64.

- `ViewExpandedText` – UTF-8 string, not more than 409600 bytes long.

  Included for Apache Hive compatibility. Not used in the normal course of AWS Glue operations.

- `TableType` – UTF-8 string, not more than 255 bytes long.

  The type of this table. AWS Glue will create tables with the `EXTERNAL_TABLE` type. Other services, such as Athena, may create tables with additional table types.

  AWS Glue related table types:

  EXTERNAL_TABLE

    Hive compatible attribute - indicates a non-Hive managed table.

  GOVERNED

    Used by AWS Lake Formation. The AWS Glue Data Catalog understands `GOVERNED`.
- `Parameters` – A map array of key-value pairs.

  Each key is a Key string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a UTF-8 string, not more than 512000 bytes long.

  These key-value pairs define properties associated with the table.
- `TargetTable` – A TableIdentifier object.

  A `TableIdentifier` structure that describes a target table for resource linking.
- `ViewDefinition` – A ViewDefinitionInput object.

  A structure that contains all the information that defines the view, including the dialect or dialects for the view, and the query.

FederatedTable structure

A table that points to an entity outside the AWS Glue Data Catalog.

Fields

- `Identifier` – UTF-8 string, not less than 1 or more than 512 bytes long, matching the Single-line string pattern.

  A unique identifier for the federated table.

- `DatabaseIdentifier` – UTF-8 string, not less than 1 or more than 512 bytes long, matching the Single-line string pattern.

  A unique identifier for the federated database.
- `ConnectionName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the connection to the external metastore.
- `ConnectionType` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The type of connection used to access the federated table, specifying the protocol or method for connecting to the external data source.

Column structure

A column in a `Table`.

Fields

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the `Column`.
- `Type` – UTF-8 string, not more than 131072 bytes long, matching the Single-line string pattern.

  The data type of the `Column`.
- `Comment` – Comment string, not more than 255 bytes long, matching the Single-line string pattern.

  A free-form text comment.
- `Parameters` – A map array of key-value pairs.

  Each key is a Key string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a UTF-8 string, not more than 512000 bytes long.

  These key-value pairs define properties associated with the column.

StorageDescriptor structure

Describes the physical storage of table data.

Fields

- `Columns` – An array of Column objects.

  A list of the `Columns` in the table.
- `Location` – Location string, not more than 2056 bytes long, matching the URI address multi-line string pattern.

  The physical location of the table. By default, this takes the form of the warehouse location, followed by the database location in the warehouse, followed by the table name.
- `AdditionalLocations` – An array of UTF-8 strings.

  A list of locations that point to the path where a Delta table is located.
- `InputFormat` – Format string, not more than 128 bytes long, matching the Single-line string pattern.

  The input format: `SequenceFileInputFormat` (binary), or `TextInputFormat`, or a custom format.
- `OutputFormat` – Format string, not more than 128 bytes long, matching the Single-line string pattern.

  The output format: `SequenceFileOutputFormat` (binary), or `IgnoreKeyTextOutputFormat`, or a custom format.
- `Compressed` – Boolean.

  `True` if the data in the table is compressed, or `False` if not.
- `NumberOfBuckets` – Number (integer).

  Must be specified if the table contains any dimension columns.
- `SerdeInfo` – A SerDeInfo object.

  The serialization/deserialization (SerDe) information.
- `BucketColumns` – An array of UTF-8 strings.

  A list of reducer grouping columns, clustering columns, and bucketing columns in the table.

- `SortColumns` – An array of [Order](#) objects.

  A list specifying the sort order of each bucket in the table.

- `Parameters` – A map array of key-value pairs.

  Each key is a Key string, not less than 1 or more than 255 bytes long, matching the [Single-line string pattern](#).

  Each value is a UTF-8 string, not more than 512000 bytes long.

  The user-supplied properties in key-value form.

- `SkewedInfo` – A [SkewedInfo](#) object.

  The information about values that appear frequently in a column (skewed values).

- `StoredAsSubDirectories` – Boolean.

  `True` if the table data is stored in subdirectories, or `False` if not.

- `SchemaReference` – A [SchemaReference](#) object.

  An object that references a schema stored in the AWS Glue Schema Registry.

  When creating a table, you can pass an empty list of columns for the schema, and instead use a schema reference.

SchemaReference structure

An object that references a schema stored in the AWS Glue Schema Registry.

Fields

- `SchemaId` – A [SchemaId](#) object.

  A structure that contains schema identity fields. Either this or the `SchemaVersionId` has to be provided.

- `SchemaVersionId` – UTF-8 string, not less than 36 or more than 36 bytes long, matching the [Custom string pattern #44](#).

  The unique ID assigned to a version of the schema. Either this or the `SchemaId` has to be provided.

- `SchemaVersionNumber` – Number (long), not less than 1 or more than 100000.

  The version number of the schema.

SerDeInfo structure

Information about a serialization/deserialization program (SerDe) that serves as an extractor and loader.

Fields

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Name of the SerDe.
- `SerializationLibrary` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Usually the class that implements the SerDe. An example is `org.apache.hadoop.hive.serde2.columnar.ColumnarSerDe`.
- `Parameters` – A map array of key-value pairs.

  Each key is a Key string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a UTF-8 string, not more than 512000 bytes long.

  These key-value pairs define initialization parameters for the SerDe.

Order structure

Specifies the sort order of a sorted column.

Fields

- `Column` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the column.
- `SortOrder` – *Required:* Number (integer), not more than 1.

Indicates that the column is sorted in ascending order (`== 1`), or in descending order (`==0`).

SkewedInfo structure

Specifies skewed values in a table. Skewed values are those that occur with very high frequency.

Fields

- `SkewedColumnNames` – An array of UTF-8 strings.

  A list of names of columns that contain skewed values.
- `SkewedColumnValues` – An array of UTF-8 strings.

  A list of values that appear so frequently as to be considered skewed.
- `SkewedColumnValueLocationMaps` – A map array of key-value pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

  A mapping of skewed values to the columns that contain them.

TableVersion structure

Specifies a version of a table.

Fields

- `Table` – A Table object.

  The table in question.
- `VersionId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID value that identifies this table version. A `VersionId` is a string representation of an integer. Each version is incremented by 1.

TableError structure

An error record for table operations.

Fields

- `TableName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table. For Hive compatibility, this must be entirely lowercase.
- `ErrorDetail` – An ErrorDetail object.

  The details about the error.

TableVersionError structure

An error record for table-version operations.

Fields

- `TableName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table in question.
- `VersionId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID value of the version in question. A `VersionID` is a string representation of an integer. Each version is incremented by 1.
- `ErrorDetail` – An ErrorDetail object.

  The details about the error.

SortCriterion structure

Specifies a field to sort by and a sort order.

Fields

- `FieldName` – Value string, not less than 1 or more than 1024 bytes long.

The name of the field on which to sort.

- `Sort` – UTF-8 string (valid values: `ASC="ASCENDING"` | `DESC="DESCENDING"`).

  An ascending or descending sort.

TableIdentifier structure

A structure that describes a target table for resource linking.

Fields

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog in which the table resides.
- `DatabaseName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the catalog database that contains the target table.
- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the target table.
- `Region` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Region of the target table.

KeySchemaElement structure

A partition key pair consisting of a name and a type.

Fields

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of a partition key.

- `Type` – *Required:* UTF-8 string, not more than 131072 bytes long, matching the Single-line string pattern.

  The type of a partition key.

PartitionIndex structure

A structure for a partition index.

Fields

- `Keys` – *Required:* An array of UTF-8 strings, at least 1 string.

  The keys for the partition index.
- `IndexName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the partition index.

PartitionIndexDescriptor structure

A descriptor for a partition index in a table.

Fields

- `IndexName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the partition index.
- `Keys` – *Required:* An array of KeySchemaElement objects, at least 1 structure.

  A list of one or more keys, as `KeySchemaElement` structures, for the partition index.
- `IndexStatus` – *Required:* UTF-8 string (valid values: `CREATING` | `ACTIVE` | `DELETING` | `FAILED`).

  The status of the partition index.

  The possible statuses are:
  - CREATING: The index is being created. When an index is in a CREATING state, the index or its table cannot be deleted.

  - ACTIVE: The index creation succeeds.
  - FAILED: The index creation fails.
  - DELETING: The index is deleted from the list of indexes.
- `BackfillErrors` – An array of [BackfillError](#) objects.

  A list of errors that can occur when registering partition indexes for an existing table.

BackfillError structure

A list of errors that can occur when registering partition indexes for an existing table.

These errors give the details about why an index registration failed and provide a limited number of partitions in the response, so that you can fix the partitions at fault and try registering the index again. The most common set of errors that can occur are categorized as follows:

- EncryptedPartitionError: The partitions are encrypted.
- InvalidPartitionTypeDataError: The partition value doesn't match the data type for that partition column.
- MissingPartitionValueError: The partitions are encrypted.
- UnsupportedPartitionCharacterError: Characters inside the partition value are not supported. For example: U+0000 , U+0001, U+0002.
- InternalError: Any error which does not belong to other error codes.

Fields

- `Code` – UTF-8 string (valid values: `ENCRYPTED_PARTITION_ERROR` | `INTERNAL_ERROR` | `INVALID_PARTITION_TYPE_DATA_ERROR` | `MISSING_PARTITION_VALUE_ERROR` | `UNSUPPORTED_PARTITION_CHARACTER_ERROR`).

  The error code for an error that occurred when registering partition indexes for an existing table.
- `Partitions` – An array of [PartitionValueList](#) objects.

  A list of a limited number of partitions in the response.

IcebergInput structure

A structure that defines an Apache Iceberg metadata table to create in the catalog.

Fields

- `MetadataOperation` – *Required:* UTF-8 string (valid values: `CREATE`).

  A required metadata operation. Can only be set to `CREATE`.
- `Version` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The table version for the Iceberg table. Defaults to 2.
- `CreateIcebergTableInput` – A CreateIcebergTableInput object.

  The configuration parameters required to create a new Iceberg table in the AWS Glue Data Catalog, including table properties and metadata specifications.

OpenTableFormatInput structure

A structure representing an open format table.

Fields

- `IcebergInput` – An IcebergInput object.

  Specifies an `IcebergInput` structure that defines an Apache Iceberg metadata table.

ViewDefinition structure

A structure containing details for representations.

Fields

- `IsProtected` – Boolean.

  You can set this flag as true to instruct the engine not to push user-provided operations into the logical plan of the view during query planning. However, setting this flag does not guarantee that the engine will comply. Refer to the engine's documentation to understand the guarantees provided, if any.

- `Definer` – UTF-8 string, not less than 20 or more than 2048 bytes long, matching the Single-line string pattern.

  The definer of a view in SQL.
- `SubObjects` – An array of UTF-8 strings, not more than 10 strings.

  A list of table Amazon Resource Names (ARNs).
- `Representations` – An array of ViewRepresentation objects, not less than 1 or more than 1000 structures.

  A list of representations.

ViewDefinitionInput structure

A structure containing details for creating or updating an AWS Glue view.

Fields

- `IsProtected` – Boolean.

  You can set this flag as true to instruct the engine not to push user-provided operations into the logical plan of the view during query planning. However, setting this flag does not guarantee that the engine will comply. Refer to the engine's documentation to understand the guarantees provided, if any.
- `Definer` – UTF-8 string, not less than 20 or more than 2048 bytes long, matching the Single-line string pattern.

  The definer of a view in SQL.
- `Representations` – An array of ViewRepresentationInput objects, not less than 1 or more than 10 structures.

  A list of structures that contains the dialect of the view, and the query that defines the view.
- `SubObjects` – An array of UTF-8 strings, not more than 10 strings.

  A list of base table ARNs that make up the view.

ViewRepresentation structure

A structure that contains the dialect of the view, and the query that defines the view.

Fields

- `Dialect` – UTF-8 string (valid values: `REDSHIFT` | `ATHENA` | `SPARK`).

  The dialect of the query engine.
- `DialectVersion` – UTF-8 string, not less than 1 or more than 255 bytes long.

  The version of the dialect of the query engine. For example, 3.0.0.
- `ViewOriginalText` – UTF-8 string, not more than 409600 bytes long.

  The `SELECT` query provided by the customer during `CREATE VIEW DDL`. This SQL is not used during a query on a view (`ViewExpandedText` is used instead). `ViewOriginalText` is used for cases like `SHOW CREATE VIEW` where users want to see the original DDL command that created the view.
- `ViewExpandedText` – UTF-8 string, not more than 409600 bytes long.

  The expanded SQL for the view. This SQL is used by engines while processing a query on a view. Engines may perform operations during view creation to transform `ViewOriginalText` to `ViewExpandedText`. For example:
  - Fully qualified identifiers: `SELECT * from table1 -> SELECT * from db1.table1`
- `ValidationConnection` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the connection to be used to validate the specific representation of the view.
- `IsStale` – Boolean.

  Dialects marked as stale are no longer valid and must be updated before they can be queried in their respective query engines.

ViewRepresentationInput structure

A structure containing details of a representation to update or create a Lake Formation view.

Fields

- `Dialect` – UTF-8 string (valid values: `REDSHIFT` | `ATHENA` | `SPARK`).

  A parameter that specifies the engine type of a specific representation.
- `DialectVersion` – UTF-8 string, not less than 1 or more than 255 bytes long.

  A parameter that specifies the version of the engine of a specific representation.
- `ViewOriginalText` – UTF-8 string, not more than 409600 bytes long.

  A string that represents the original SQL query that describes the view.
- `ValidationConnection` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the connection to be used to validate the specific representation of the view.
- `ViewExpandedText` – UTF-8 string, not more than 409600 bytes long.

  A string that represents the SQL query that describes the view with expanded resource ARNs

UpdateOpenTableFormatInput structure

Input parameters for updating open table format tables in AWS GlueData Catalog, serving as a wrapper for format-specific update operations such as Apache Iceberg.

Fields

- `UpdateIcebergInput` – An UpdateIcebergInput object.

  Apache Iceberg-specific update parameters that define the table modifications to be applied, including schema changes, partition specifications, and table properties.

UpdateIcebergInput structure

Input parameters specific to updating Apache Iceberg tables in AWS Glue Data Catalog, containing the update operations to be applied to an existing Iceberg table.

Fields

- `UpdateIcebergTableInput` – *Required:* An UpdateIcebergTableInput object.

The specific update operations to be applied to the Iceberg table, containing a list of updates that define the new state of the table including schema, partitions, and properties.

CreateIcebergTableInput structure

The configuration parameters required to create a new Iceberg table in the AWS Glue Data Catalog, including table properties and metadata specifications.

Fields

- `Location` – *Required:* Location string, not more than 2056 bytes long, matching the URI address multi-line string pattern.

  The S3 location where the Iceberg table data will be stored.
- `Schema` – *Required:* An IcebergSchema object.

  The schema definition that specifies the structure, field types, and metadata for the Iceberg table.
- `PartitionSpec` – An IcebergPartitionSpec object.

  The partitioning specification that defines how the Iceberg table data will be organized and partitioned for optimal query performance.
- `WriteOrder` – An IcebergSortOrder object.

  The sort order specification that defines how data should be ordered within each partition to optimize query performance.
- `Properties` – A map array of key-value pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

  Key-value pairs of additional table properties and configuration settings for the Iceberg table.

UpdateIcebergTableInput structure

Contains the update operations to be applied to an existing Iceberg table inAWS Glue Data Catalog, defining the new state of the table metadata.

Fields

- `Updates` – *Required:* An array of IcebergTableUpdate objects.

  The list of table update operations that specify the changes to be made to the Iceberg table, including schema modifications, partition specifications, and table properties.

IcebergSortOrder structure

Defines the sort order specification for an Iceberg table, determining how data should be ordered within partitions to optimize query performance.

Fields

- `OrderId` – *Required:* Number (integer).

  The unique identifier for this sort order specification within the Iceberg table's metadata.
- `Fields` – *Required:* An array of IcebergSortField objects.

  The list of fields and their sort directions that define the ordering criteria for the Iceberg table data.

IcebergSortField structure

Defines a single field within an Iceberg sort order specification, including the source field, transformation, sort direction, and null value ordering.

Fields

- `SourceId` – *Required:* Number (integer).

  The identifier of the source field from the table schema that this sort field is based on.
- `Transform` – *Required:* UTF-8 string.

  The transformation function applied to the source field before sorting, such as identity, bucket, or truncate.
- `Direction` – *Required:* UTF-8 string (valid values: `asc="ASC"` | `desc="DESC"`).

  The sort direction for this field, either ascending or descending.

- `NullOrder` – *Required:* UTF-8 string (valid values: `nulls-first="NULLS_FIRST"` | `nulls-last="NULLS_LAST"`).

  The ordering behavior for null values in this field, specifying whether nulls should appear first or last in the sort order.

IcebergPartitionSpec structure

Defines the partitioning specification for an Iceberg table, determining how table data will be organized and partitioned for optimal query performance.

Fields

- `Fields` – *Required:* An array of IcebergPartitionField objects.

  The list of partition fields that define how the table data should be partitioned, including source fields and their transformations.
- `SpecId` – Number (integer).

  The unique identifier for this partition specification within the Iceberg table's metadata history.

IcebergPartitionField structure

Defines a single partition field within an Iceberg partition specification, including the source field, transformation function, partition name, and unique identifier.

Fields

- `SourceId` – *Required:* Number (integer).

  The identifier of the source field from the table schema that this partition field is based on.
- `Transform` – *Required:* UTF-8 string.

  The transformation function applied to the source field to create the partition, such as identity, bucket, truncate, year, month, day, or hour.
- `Name` – *Required:* UTF-8 string, not less than 1 or more than 1024 bytes long, matching the Single-line string pattern.

  The name of the partition field as it will appear in the partitioned table structure.

- `FieldId` – Number (integer).

  The unique identifier assigned to this partition field within the Iceberg table's partition specification.

IcebergSchema structure

Defines the schema structure for an Iceberg table, including field definitions, data types, and schema metadata.

Fields

- `SchemaId` – Number (integer).

  The unique identifier for this schema version within the Iceberg table's schema evolution history.
- `IdentifierFieldIds` – An array of signed 32-bit integers.

  The list of field identifiers that uniquely identify records in the table, used for row-level operations and deduplication.
- `Type` – UTF-8 string (valid values: `struct="STRUCT"`).

  The root type of the schema structure, typically "struct" for Iceberg table schemas.
- `Fields` – *Required:* An array of IcebergStructField objects.

  The list of field definitions that make up the table schema, including field names, types, and metadata.

IcebergStructField structure

Defines a single field within an Iceberg table schema, including its identifier, name, data type, nullability, and documentation.

Fields

- `Id` – *Required:* Number (integer).

  The unique identifier assigned to this field within the Iceberg table schema, used for schema evolution and field tracking.

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 1024 bytes long, matching the Single-line string pattern.

  The name of the field as it appears in the table schema and query operations.
- `Type` – *Required:* An empty-structure named `IcebergDocument`.

  The data type definition for this field, specifying the structure and format of the data it contains.
- `Required` – *Required:* Boolean.

  Indicates whether this field is required (non-nullable) or optional (nullable) in the table schema.
- `Doc` – Comment string, not more than 255 bytes long, matching the Single-line string pattern.

  Optional documentation or description text that provides additional context about the purpose and usage of this field.

IcebergTableUpdate structure

Defines a complete set of updates to be applied to an Iceberg table, including schema changes, partitioning modifications, sort order adjustments, location updates, and property changes.

Fields

- `Schema` – *Required:* An IcebergSchema object.

  The updated schema definition for the Iceberg table, specifying any changes to field structure, data types, or schema metadata.
- `PartitionSpec` – An IcebergPartitionSpec object.

  The updated partitioning specification that defines how the table data should be reorganized and partitioned.
- `SortOrder` – An IcebergSortOrder object.

  The updated sort order specification that defines how data should be ordered within partitions for optimal query performance.
- `Location` – *Required:* Location string, not more than 2056 bytes long, matching the URI address multi-line string pattern.

  The updated S3 location where the Iceberg table data will be stored.

- `Properties` – A map array of key-value pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

  Updated key-value pairs of table properties and configuration settings for the Iceberg table.

Operations

- CreateTable action (Python: create_table)
- UpdateTable action (Python: update_table)
- DeleteTable action (Python: delete_table)
- BatchDeleteTable action (Python: batch_delete_table)
- GetTable action (Python: get_table)
- GetTables action (Python: get_tables)
- GetTableVersion action (Python: get_table_version)
- GetTableVersions action (Python: get_table_versions)
- DeleteTableVersion action (Python: delete_table_version)
- BatchDeleteTableVersion action (Python: batch_delete_table_version)
- SearchTables action (Python: search_tables)
- GetPartitionIndexes action (Python: get_partition_indexes)
- CreatePartitionIndex action (Python: create_partition_index)
- DeletePartitionIndex action (Python: delete_partition_index)
- GetColumnStatisticsForTable action (Python: get_column_statistics_for_table)
- UpdateColumnStatisticsForTable action (Python: update_column_statistics_for_table)
- DeleteColumnStatisticsForTable action (Python: delete_column_statistics_for_table)

CreateTable action (Python: create_table)

Creates a new table definition in the Data Catalog.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog in which to create the `Table`. If none is supplied, the AWS account ID is used by default.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The catalog database in which to create the new table. For Hive compatibility, this name is entirely lowercase.
- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique identifier for the table within the specified database that will be created in the AWS Glue Data Catalog.
- `TableInput` – A TableInput object.

  The `TableInput` object that defines the metadata table to create in the catalog.
- `PartitionIndexes` – An array of PartitionIndex objects, not more than 3 structures.

  A list of partition indexes, `PartitionIndex` structures, to create in the table.
- `TransactionId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #43.

  The ID of the transaction.
- `OpenTableFormatInput` – An OpenTableFormatInput object.

  Specifies an `OpenTableFormatInput` structure when creating an open format table.

Response

- *No Response parameters.*

Errors

- `AlreadyExistsException`

- `InvalidInputException`
- `EntityNotFoundException`
- `ResourceNumberLimitExceededException`
- `InternalServiceException`
- `OperationTimeoutException`
- `GlueEncryptionException`
- `ConcurrentModificationException`
- `ResourceNotReadyException`
- `FederationSourceException`
- `FederationSourceRetryableException`

UpdateTable action (Python: update_table)

Updates a metadata table in the Data Catalog.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog where the table resides. If none is provided, the AWS account ID is used by default.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the catalog database in which the table resides. For Hive compatibility, this name is entirely lowercase.
- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique identifier for the table within the specified database that will be created in the AWS Glue Data Catalog.
- `TableInput` – A TableInput object.

  An updated `TableInput` object to define the metadata table in the catalog.

- `SkipArchive` – Boolean.

  By default, `UpdateTable` always creates an archived version of the table before updating it. However, if `skipArchive` is set to true, `UpdateTable` does not create the archived version.
- `TransactionId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #43.

  The transaction ID at which to update the table contents.
- `VersionId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The version ID at which to update the table contents.
- `ViewUpdateAction` – UTF-8 string (valid values: `ADD` | `REPLACE` | `ADD_OR_REPLACE` | `DROP`).

  The operation to be performed when updating the view.
- `Force` – Boolean.

  A flag that can be set to true to ignore matching storage descriptor and subobject matching requirements.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `ConcurrentModificationException`
- `ResourceNumberLimitExceededException`
- `GlueEncryptionException`
- `ResourceNotReadyException`
- `FederationSourceException`

- `FederationSourceRetryableException`
- `AlreadyExistsException`

DeleteTable action (Python: delete_table)

Removes a table definition from the Data Catalog.

> **Note**
>
> After completing this operation, you no longer have access to the table versions and partitions that belong to the deleted table. AWS Glue deletes these "orphaned" resources asynchronously in a timely manner, at the discretion of the service.
> To ensure the immediate deletion of all related resources, before calling `DeleteTable`, use `DeleteTableVersion` or `BatchDeleteTableVersion`, and `DeletePartition` or `BatchDeletePartition`, to delete any resources that belong to the table.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog where the table resides. If none is provided, the AWS account ID is used by default.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the catalog database in which the table resides. For Hive compatibility, this name is entirely lowercase.
- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table to be deleted. For Hive compatibility, this name is entirely lowercase.
- `TransactionId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #43.

  The transaction ID at which to delete the table contents.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `ConcurrentModificationException`
- `ResourceNotReadyException`
- `FederationSourceException`
- `FederationSourceRetryableException`

BatchDeleteTable action (Python: batch_delete_table)

Deletes multiple tables at once.

> **Note**
>
> After completing this operation, you no longer have access to the table versions and partitions that belong to the deleted table. AWS Glue deletes these "orphaned" resources asynchronously in a timely manner, at the discretion of the service.
> To ensure the immediate deletion of all related resources, before calling `BatchDeleteTable`, use `DeleteTableVersion` or `BatchDeleteTableVersion`, and `DeletePartition` or `BatchDeletePartition`, to delete any resources that belong to the table.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog where the table resides. If none is provided, the AWS account ID is used by default.

- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the catalog database in which the tables to delete reside. For Hive compatibility, this name is entirely lowercase.
- `TablesToDelete` – *Required:* An array of UTF-8 strings, not more than 100 strings.

  A list of the table to delete.
- `TransactionId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #43.

  The transaction ID at which to delete the table contents.

Response

- `Errors` – An array of TableError objects.

  A list of errors encountered in attempting to delete the specified tables.

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`
- `GlueEncryptionException`
- `ResourceNotReadyException`

GetTable action (Python: get_table)

Retrieves the `Table` definition in a Data Catalog for a specified table.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The ID of the Data Catalog where the table resides. If none is provided, the AWS account ID is used by default.

- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the database in the catalog in which the table resides. For Hive compatibility, this name is entirely lowercase.

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table for which to retrieve the definition. For Hive compatibility, this name is entirely lowercase.

- `TransactionId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #43.

  The transaction ID at which to read the table contents.

- `QueryAsOfTime` – Timestamp.

  The time as of when to read the table contents. If not set, the most recent transaction commit time will be used. Cannot be specified along with `TransactionId`.

- `IncludeStatusDetails` – Boolean.

  Specifies whether to include status details related to a request to create or update an AWS Glue Data Catalog view.

Response

- `Table` – A Table object.

  The `Table` object that defines the specified table.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`

- `OperationTimeoutException`
- `GlueEncryptionException`
- `ResourceNotReadyException`
- `FederationSourceException`
- `FederationSourceRetryableException`

GetTables action (Python: get_tables)

Retrieves the definitions of some or all of the tables in a given `Database`.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog where the tables reside. If none is provided, the AWS account ID is used by default.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The database in the catalog whose tables to list. For Hive compatibility, this name is entirely lowercase.
- `Expression` – UTF-8 string, not more than 2048 bytes long, matching the Single-line string pattern.

  A regular expression pattern. If present, only those tables whose names match the pattern are returned.
- `NextToken` – UTF-8 string.

  A continuation token, included if this is a continuation call.
- `MaxResults` – Number (integer), not less than 1 or more than 100.

  The maximum number of tables to return in a single response.
- `TransactionId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #43.

  The transaction ID at which to read the table contents.

- `QueryAsOfTime` – Timestamp.

  The time as of when to read the table contents. If not set, the most recent transaction commit time will be used. Cannot be specified along with `TransactionId`.
- `IncludeStatusDetails` – Boolean.

  Specifies whether to include status details related to a request to create or update an AWS Glue Data Catalog view.
- `AttributesToGet` – An array of UTF-8 strings.

  Specifies the table fields returned by the `GetTables` call. This parameter doesn't accept an empty list. The request must include `NAME`.

  The following are the valid combinations of values:
  - `NAME` - Names of all tables in the database.
  - `NAME`, `TABLE_TYPE` - Names of all tables and the table types.

Response

- `TableList` – An array of Table objects.

  A list of the requested `Table` objects.
- `NextToken` – UTF-8 string.

  A continuation token, present if the current list segment is not the last.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`
- `GlueEncryptionException`
- `FederationSourceException`
- `FederationSourceRetryableException`

GetTableVersion action (Python: get_table_version)

Retrieves a specified version of a table.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog where the tables reside. If none is provided, the AWS account ID is used by default.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The database in the catalog in which the table resides. For Hive compatibility, this name is entirely lowercase.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table. For Hive compatibility, this name is entirely lowercase.
- `VersionId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID value of the table version to be retrieved. A `VersionID` is a string representation of an integer. Each version is incremented by 1.

Response

- `TableVersion` – A TableVersion object.

  The requested table version.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`

- `GlueEncryptionException`

GetTableVersions action (Python: get_table_versions)

Retrieves a list of strings that identify available versions of a specified table.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog where the tables reside. If none is provided, the AWS account ID is used by default.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The database in the catalog in which the table resides. For Hive compatibility, this name is entirely lowercase.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table. For Hive compatibility, this name is entirely lowercase.
- `NextToken` – UTF-8 string.

  A continuation token, if this is not the first call.
- `MaxResults` – Number (integer), not less than 1 or more than 100.

  The maximum number of table versions to return in one response.

Response

- `TableVersions` – An array of TableVersion objects.

  A list of strings identifying available versions of the specified table.
- `NextToken` – UTF-8 string.

  A continuation token, if the list of available versions does not include the last one.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `GlueEncryptionException`

DeleteTableVersion action (Python: delete_table_version)

Deletes a specified version of a table.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog where the tables reside. If none is provided, the AWS account ID is used by default.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The database in the catalog in which the table resides. For Hive compatibility, this name is entirely lowercase.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table. For Hive compatibility, this name is entirely lowercase.
- `VersionId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the table version to be deleted. A `VersionID` is a string representation of an integer. Each version is incremented by 1.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`

BatchDeleteTableVersion action (Python: batch_delete_table_version)

Deletes a specified batch of versions of a table.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog where the tables reside. If none is provided, the AWS account ID is used by default.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The database in the catalog in which the table resides. For Hive compatibility, this name is entirely lowercase.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table. For Hive compatibility, this name is entirely lowercase.
- `VersionIds` – *Required:* An array of UTF-8 strings, not more than 100 strings.

  A list of the IDs of versions to be deleted. A `VersionId` is a string representation of an integer. Each version is incremented by 1.

Response

- `Errors` – An array of TableVersionError objects.

  A list of errors encountered while trying to delete the specified table versions.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`

SearchTables action (Python: search_tables)

Searches a set of tables based on properties in the table metadata as well as on the parent database. You can search against text or filter conditions.

You can only get tables that you have access to based on the security policies defined in Lake Formation. You need at least a read-only access to the table for it to be returned. If you do not have access to all the columns in the table, these columns will not be searched against when returning the list of tables back to you. If you have access to the columns but not the data in the columns, those columns and the associated metadata for those columns will be included in the search.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A unique identifier, consisting of *`account_id`*.
- `NextToken` – UTF-8 string.

  A continuation token, included if this is a continuation call.
- `Filters` – An array of PropertyPredicate objects.

  A list of key-value pairs, and a comparator used to filter the search results. Returns all entities matching the predicate.

  The `Comparator` member of the `PropertyPredicate` struct is used only for time fields, and can be omitted for other field types. Also, when comparing string values, such as when `Key=Name`, a fuzzy match algorithm is used. The `Key` field (for example, the value of the `Name` field) is split on certain punctuation characters, for example, -, :, #, etc. into tokens. Then each token is exact-match compared with the `Value` member of `PropertyPredicate`. For example, if `Key=Name` and `Value=link`, tables named `customer-link` and `xx-link-yy` are returned, but `xxlinkyy` is not returned.

- `SearchText` – Value string, not less than 1 or more than 1024 bytes long.

  A string used for a text search.

  Specifying a value in quotes filters based on an exact match to the value.
- `SortCriteria` – An array of SortCriterion objects, not more than 1 structures.

  A list of criteria for sorting the results by a field name, in an ascending or descending order.
- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum number of tables to return in a single response.
- `ResourceShareType` – UTF-8 string (valid values: `FOREIGN` | `ALL` | `FEDERATED`).

  Allows you to specify that you want to search the tables shared with your account. The allowable values are `FOREIGN` or `ALL`.
  - If set to `FOREIGN`, will search the tables shared with your account.
  - If set to `ALL`, will search the tables shared with your account, as well as the tables in yor local account.
- `IncludeStatusDetails` – Boolean.

  Specifies whether to include status details related to a request to create or update an AWS Glue Data Catalog view.

Response

- `NextToken` – UTF-8 string.

  A continuation token, present if the current list segment is not the last.
- `TableList` – An array of Table objects.

  A list of the requested `Table` objects. The `SearchTables` response returns only the tables that you have access to.

Errors

- `InternalServiceException`
- `InvalidInputException`
- `OperationTimeoutException`

GetPartitionIndexes action (Python: get_partition_indexes)

Retrieves the partition indexes associated with a table.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The catalog ID where the table resides.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Specifies the name of a database from which you want to retrieve partition indexes.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Specifies the name of a table for which you want to retrieve the partition indexes.
- `NextToken` – UTF-8 string.

  A continuation token, included if this is a continuation call.

Response

- `PartitionIndexDescriptorList` – An array of PartitionIndexDescriptor objects.

  A list of index descriptors.
- `NextToken` – UTF-8 string.

  A continuation token, present if the current list segment is not the last.

Errors

- `InternalServiceException`
- `OperationTimeoutException`
- `InvalidInputException`
- `EntityNotFoundException`
- `ConflictException`

CreatePartitionIndex action (Python: create_partition_index)

Creates a specified partition index in an existing table.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The catalog ID where the table resides.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Specifies the name of a database in which you want to create a partition index.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Specifies the name of a table in which you want to create a partition index.
- `PartitionIndex` – *Required:* A PartitionIndex object.

  Specifies a `PartitionIndex` structure to create a partition index in an existing table.

Response

- *No Response parameters.*

Errors

- `AlreadyExistsException`

- `InvalidInputException`
- `EntityNotFoundException`
- `ResourceNumberLimitExceededException`
- `InternalServiceException`
- `OperationTimeoutException`
- `GlueEncryptionException`

DeletePartitionIndex action (Python: delete_partition_index)

Deletes a specified partition index from an existing table.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The catalog ID where the table resides.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Specifies the name of a database from which you want to delete a partition index.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Specifies the name of a table from which you want to delete a partition index.
- `IndexName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the partition index to be deleted.

Response

- *No Response parameters.*

Errors

- `InternalServiceException`

- `OperationTimeoutException`
- `InvalidInputException`
- `EntityNotFoundException`
- `ConflictException`
- `GlueEncryptionException`

GetColumnStatisticsForTable action (Python: get_column_statistics_for_table)

Retrieves table statistics of columns.

The Identity and Access Management (IAM) permission required for this operation is `GetTable`.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog where the partitions in question reside. If none is supplied, the AWS account ID is used by default.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the catalog database where the partitions reside.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the partitions' table.
- `ColumnNames` – *Required:* An array of UTF-8 strings, not more than 100 strings.

  A list of the column names.

Response

- `ColumnStatisticsList` – An array of ColumnStatistics objects.

  List of ColumnStatistics.
- `Errors` – An array of ColumnError objects.

List of ColumnStatistics that failed to be retrieved.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `GlueEncryptionException`

UpdateColumnStatisticsForTable action (Python: update_column_statistics_for_table)

Creates or updates table statistics of columns.

The Identity and Access Management (IAM) permission required for this operation is `UpdateTable`.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog where the partitions in question reside. If none is supplied, the AWS account ID is used by default.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the catalog database where the partitions reside.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the partitions' table.
- `ColumnStatisticsList` – *Required:* An array of ColumnStatistics objects, not more than 25 structures.

  A list of the column statistics.

Response

- `Errors` – An array of ColumnStatisticsError objects.

  List of ColumnStatisticsErrors.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `GlueEncryptionException`

DeleteColumnStatisticsForTable action (Python: delete_column_statistics_for_table)

Retrieves table statistics of columns.

The Identity and Access Management (IAM) permission required for this operation is `DeleteTable`.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog where the partitions in question reside. If none is supplied, the AWS account ID is used by default.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the catalog database where the partitions reside.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the partitions' table.

- `ColumnName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the column.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `GlueEncryptionException`

Partition API

The Partition API describes data types and operations used to work with partitions.

Data types

- Partition structure
- PartitionInput structure
- PartitionSpecWithSharedStorageDescriptor structure
- PartitionListComposingSpec structure
- PartitionSpecProxy structure
- PartitionValueList structure
- Segment structure
- PartitionError structure
- BatchUpdatePartitionFailureEntry structure
- BatchUpdatePartitionRequestEntry structure

- StorageDescriptor structure
- SchemaReference structure
- SerDeInfo structure
- SkewedInfo structure

Partition structure

Represents a slice of table data.

Fields

- `Values` – An array of UTF-8 strings.

  The values of the partition.
- `DatabaseName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the catalog database in which to create the partition.
- `TableName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the database table in which to create the partition.
- `CreationTime` – Timestamp.

  The time at which the partition was created.
- `LastAccessTime` – Timestamp.

  The last time at which the partition was accessed.
- `StorageDescriptor` – A StorageDescriptor object.

  Provides information about the physical location where the partition is stored.
- `Parameters` – A map array of key-value pairs.

  Each key is a Key string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a UTF-8 string, not more than 512000 bytes long.

These key-value pairs define partition parameters.

- `LastAnalyzedTime` – Timestamp.

  The last time at which column statistics were computed for this partition.

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog in which the partition resides.

PartitionInput structure

The structure used to create and update a partition.

Fields

- `Values` – An array of UTF-8 strings.

  The values of the partition. Although this parameter is not required by the SDK, you must specify this parameter for a valid input.

  The values for the keys for the new partition must be passed as an array of String objects that must be ordered in the same order as the partition keys appearing in the Amazon S3 prefix. Otherwise AWS Glue will add the values to the wrong keys.

- `LastAccessTime` – Timestamp.

  The last time at which the partition was accessed.

- `StorageDescriptor` – A StorageDescriptor object.

  Provides information about the physical location where the partition is stored.

- `Parameters` – A map array of key-value pairs.

  Each key is a Key string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a UTF-8 string, not more than 512000 bytes long.

  These key-value pairs define partition parameters.

- `LastAnalyzedTime` – Timestamp.

The last time at which column statistics were computed for this partition.

PartitionSpecWithSharedStorageDescriptor structure

A partition specification for partitions that share a physical location.

Fields

- `StorageDescriptor` – A StorageDescriptor object.

  The shared physical storage information.
- `Partitions` – An array of Partition objects.

  A list of the partitions that share this physical location.

PartitionListComposingSpec structure

Lists the related partitions.

Fields

- `Partitions` – An array of Partition objects.

  A list of the partitions in the composing specification.

PartitionSpecProxy structure

Provides a root path to specified partitions.

Fields

- `DatabaseName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The catalog database in which the partitions reside.
- `TableName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table that contains the partitions.

- `RootPath` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The root path of the proxy for addressing the partitions.
- `PartitionSpecWithSharedSD` – A PartitionSpecWithSharedStorageDescriptor object.

  A specification of partitions that share the same physical storage location.
- `PartitionListComposingSpec` – A PartitionListComposingSpec object.

  Specifies a list of partitions.

PartitionValueList structure

Contains a list of values defining partitions.

Fields

- `Values` – *Required:* An array of UTF-8 strings.

  The list of values.

Segment structure

Defines a non-overlapping region of a table's partitions, allowing multiple requests to be run in parallel.

Fields

- `SegmentNumber` – *Required:* Number (integer), not more than None.

  The zero-based index number of the segment. For example, if the total number of segments is 4, `SegmentNumber` values range from 0 through 3.
- `TotalSegments` – *Required:* Number (integer), not less than 1 or more than 10.

  The total number of segments.

PartitionError structure

Contains information about a partition error.

Fields

- `PartitionValues` – An array of UTF-8 strings.

  The values that define the partition.
- `ErrorDetail` – An ErrorDetail object.

  The details about the partition error.

BatchUpdatePartitionFailureEntry structure

Contains information about a batch update partition error.

Fields

- `PartitionValueList` – An array of UTF-8 strings, not more than 100 strings.

  A list of values defining the partitions.
- `ErrorDetail` – An ErrorDetail object.

  The details about the batch update partition error.

BatchUpdatePartitionRequestEntry structure

A structure that contains the values and structure used to update a partition.

Fields

- `PartitionValueList` – *Required:* An array of UTF-8 strings, not more than 100 strings.

  A list of values defining the partitions.
- `PartitionInput` – *Required:* A PartitionInput object.

  The structure used to update a partition.

StorageDescriptor structure

Describes the physical storage of table data.

Fields

- `Columns` – An array of Column objects.

  A list of the `Columns` in the table.
- `Location` – Location string, not more than 2056 bytes long, matching the URI address multi-line string pattern.

  The physical location of the table. By default, this takes the form of the warehouse location, followed by the database location in the warehouse, followed by the table name.
- `AdditionalLocations` – An array of UTF-8 strings.

  A list of locations that point to the path where a Delta table is located.
- `InputFormat` – Format string, not more than 128 bytes long, matching the Single-line string pattern.

  The input format: `SequenceFileInputFormat` (binary), or `TextInputFormat`, or a custom format.
- `OutputFormat` – Format string, not more than 128 bytes long, matching the Single-line string pattern.

  The output format: `SequenceFileOutputFormat` (binary), or `IgnoreKeyTextOutputFormat`, or a custom format.
- `Compressed` – Boolean.

  `True` if the data in the table is compressed, or `False` if not.
- `NumberOfBuckets` – Number (integer).

  Must be specified if the table contains any dimension columns.
- `SerdeInfo` – A SerDeInfo object.

  The serialization/deserialization (SerDe) information.
- `BucketColumns` – An array of UTF-8 strings.

  A list of reducer grouping columns, clustering columns, and bucketing columns in the table.
- `SortColumns` – An array of Order objects.

  A list specifying the sort order of each bucket in the table.

- `Parameters` – A map array of key-value pairs.

  Each key is a Key string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a UTF-8 string, not more than 512000 bytes long.

  The user-supplied properties in key-value form.
- `SkewedInfo` – A SkewedInfo object.

  The information about values that appear frequently in a column (skewed values).
- `StoredAsSubDirectories` – Boolean.

  `True` if the table data is stored in subdirectories, or `False` if not.
- `SchemaReference` – A SchemaReference object.

  An object that references a schema stored in the AWS Glue Schema Registry.

  When creating a table, you can pass an empty list of columns for the schema, and instead use a schema reference.

SchemaReference structure

An object that references a schema stored in the AWS Glue Schema Registry.

Fields

- `SchemaId` – A SchemaId object.

  A structure that contains schema identity fields. Either this or the `SchemaVersionId` has to be provided.
- `SchemaVersionId` – UTF-8 string, not less than 36 or more than 36 bytes long, matching the Custom string pattern #44.

  The unique ID assigned to a version of the schema. Either this or the `SchemaId` has to be provided.
- `SchemaVersionNumber` – Number (long), not less than 1 or more than 100000.

  The version number of the schema.

SerDeInfo structure

Information about a serialization/deserialization program (SerDe) that serves as an extractor and loader.

Fields

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Name of the SerDe.
- `SerializationLibrary` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Usually the class that implements the SerDe. An example is `org.apache.hadoop.hive.serde2.columnar.ColumnarSerDe`.
- `Parameters` – A map array of key-value pairs.

  Each key is a Key string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a UTF-8 string, not more than 512000 bytes long.

  These key-value pairs define initialization parameters for the SerDe.

SkewedInfo structure

Specifies skewed values in a table. Skewed values are those that occur with very high frequency.

Fields

- `SkewedColumnNames` – An array of UTF-8 strings.

  A list of names of columns that contain skewed values.
- `SkewedColumnValues` – An array of UTF-8 strings.

  A list of values that appear so frequently as to be considered skewed.
- `SkewedColumnValueLocationMaps` – A map array of key-value pairs.

  Each key is a UTF-8 string.

Each value is a UTF-8 string.

A mapping of skewed values to the columns that contain them.

Operations

- CreatePartition action (Python: create_partition)
- BatchCreatePartition action (Python: batch_create_partition)
- UpdatePartition action (Python: update_partition)
- DeletePartition action (Python: delete_partition)
- BatchDeletePartition action (Python: batch_delete_partition)
- GetPartition action (Python: get_partition)
- GetPartitions action (Python: get_partitions)
- BatchGetPartition action (Python: batch_get_partition)
- BatchUpdatePartition action (Python: batch_update_partition)
- GetColumnStatisticsForPartition action (Python: get_column_statistics_for_partition)
- UpdateColumnStatisticsForPartition action (Python: update_column_statistics_for_partition)
- DeleteColumnStatisticsForPartition action (Python: delete_column_statistics_for_partition)

CreatePartition action (Python: create_partition)

Creates a new partition.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The AWS account ID of the catalog in which the partition is to be created.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the metadata database in which the partition is to be created.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The name of the metadata table in which the partition is to be created.

- `PartitionInput` – *Required:* A PartitionInput object.

  A `PartitionInput` structure defining the partition to be created.

Response

- *No Response parameters.*

Errors

- `InvalidInputException`
- `AlreadyExistsException`
- `ResourceNumberLimitExceededException`
- `InternalServiceException`
- `EntityNotFoundException`
- `OperationTimeoutException`
- `GlueEncryptionException`

BatchCreatePartition action (Python: batch_create_partition)

Creates one or more partitions in a batch operation.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the catalog in which the partition is to be created. Currently, this should be the AWS account ID.

- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the metadata database in which the partition is to be created.

- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The name of the metadata table in which the partition is to be created.

- `PartitionInputList` – *Required:* An array of PartitionInput objects, not more than 100 structures.

  A list of `PartitionInput` structures that define the partitions to be created.

Response

- `Errors` – An array of PartitionError objects.

  The errors encountered when trying to create the requested partitions.

Errors

- `InvalidInputException`
- `AlreadyExistsException`
- `ResourceNumberLimitExceededException`
- `InternalServiceException`
- `EntityNotFoundException`
- `OperationTimeoutException`
- `GlueEncryptionException`

UpdatePartition action (Python: update_partition)

Updates a partition.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog where the partition to be updated resides. If none is provided, the AWS account ID is used by default.

- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The name of the catalog database in which the table in question resides.

- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table in which the partition to be updated is located.

- `PartitionValueList` – *Required:* An array of UTF-8 strings, not more than 100 strings.

  List of partition key values that define the partition to update.

- `PartitionInput` – *Required:* A PartitionInput object.

  The new partition object to update the partition to.

  The `Values` property can't be changed. If you want to change the partition key values for a partition, delete and recreate the partition.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `GlueEncryptionException`

DeletePartition action (Python: delete_partition)

Deletes a specified partition.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The ID of the Data Catalog where the partition to be deleted resides. If none is provided, the AWS account ID is used by default.

- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the catalog database in which the table in question resides.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table that contains the partition to be deleted.
- `PartitionValues` – *Required:* An array of UTF-8 strings.

  The values that define the partition.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`

BatchDeletePartition action (Python: batch_delete_partition)

Deletes one or more partitions in a batch operation.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog where the partition to be deleted resides. If none is provided, the AWS account ID is used by default.

- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the catalog database in which the table in question resides.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table that contains the partitions to be deleted.
- `PartitionsToDelete` – *Required:* An array of PartitionValueList objects, not more than 25 structures.

  A list of `PartitionInput` structures that define the partitions to be deleted.

Response

- `Errors` – An array of PartitionError objects.

  The errors encountered when trying to delete the requested partitions.

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`

GetPartition action (Python: get_partition)

Retrieves information about a specified partition.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog where the partition in question resides. If none is provided, the AWS account ID is used by default.

- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the catalog database where the partition resides.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the partition's table.
- `PartitionValues` – *Required:* An array of UTF-8 strings.

  The values that define the partition.

Response

- `Partition` – A Partition object.

  The requested information, in the form of a `Partition` object.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `GlueEncryptionException`
- `FederationSourceException`
- `FederationSourceRetryableException`

GetPartitions action (Python: get_partitions)

Retrieves information about the partitions in a table.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The ID of the Data Catalog where the partitions in question reside. If none is provided, the AWS account ID is used by default.

- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the catalog database where the partitions reside.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the partitions' table.
- `Expression` – Predicate string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  An expression that filters the partitions to be returned.

  The expression uses SQL syntax similar to the SQL `WHERE` filter clause. The SQL statement parser JSQLParser parses the expression.

  Operators: The following are the operators that you can use in the `Expression` API call:

  =

  Checks whether the values of the two operands are equal; if yes, then the condition becomes true.

  Example: Assume 'variable a' holds 10 and 'variable b' holds 20.

  (a = b) is not true.

  < >

  Checks whether the values of two operands are equal; if the values are not equal, then the condition becomes true.

  Example: (a < > b) is true.

  >

  Checks whether the value of the left operand is greater than the value of the right operand; if yes, then the condition becomes true.

  Example: (a > b) is not true.

`<`

Checks whether the value of the left operand is less than the value of the right operand; if yes, then the condition becomes true.

Example: (a < b) is true.

`>=`

Checks whether the value of the left operand is greater than or equal to the value of the right operand; if yes, then the condition becomes true.

Example: (a >= b) is not true.

`<=`

Checks whether the value of the left operand is less than or equal to the value of the right operand; if yes, then the condition becomes true.

Example: (a <= b) is true.

AND, OR, IN, BETWEEN, LIKE, NOT, IS NULL

Logical operators.

Supported Partition Key Types: The following are the supported partition keys.

- `string`
- `date`
- `timestamp`
- `int`
- `bigint`
- `long`
- `tinyint`
- `smallint`
- `decimal`

If an type is encountered that is not valid, an exception is thrown.

The following list shows the valid operators on each type. When you define a crawler, the `partitionKey` type is created as a `STRING`, to be compatible with the catalog partitions.

Sample API Call:

Example

The table `twitter_partition` has three partitions:

```
year = 2015
        year = 2016
        year = 2017
```

Example

Get partition `year` equal to 2015

```
aws glue get-partitions --database-name dbname --table-name twitter_partition
        --expression "year*=*'2015'"
```

Example

Get partition `year` between 2016 and 2018 (exclusive)

```
aws glue get-partitions --database-name dbname --table-name twitter_partition
        --expression "year>'2016' AND year<'2018'"
```

Example

Get partition `year` between 2015 and 2018 (inclusive). The following API calls are equivalent to each other:

```
aws glue get-partitions --database-name dbname --table-name twitter_partition
        --expression "year>='2015' AND year<='2018'"

        aws glue get-partitions --database-name dbname --table-name
 twitter_partition
        --expression "year BETWEEN 2015 AND 2018"

        aws glue get-partitions --database-name dbname --table-name
 twitter_partition
        --expression "year IN (2015,2016,2017,2018)"
```

Example

A wildcard partition filter, where the following call output is partition year=2017. A regular expression is not supported in `LIKE`.

```
aws glue get-partitions --database-name dbname --table-name twitter_partition
        --expression "year LIKE '%7'"
```

- `NextToken` – UTF-8 string.

  A continuation token, if this is not the first call to retrieve these partitions.
- `Segment` – A Segment object.

  The segment of the table's partitions to scan in this request.
- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum number of partitions to return in a single response.
- `ExcludeColumnSchema` – Boolean.

  When true, specifies not returning the partition column schema. Useful when you are interested only in other partition attributes such as partition values or location. This approach avoids the problem of a large response by not returning duplicate data.
- `TransactionId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #43.

  The transaction ID at which to read the partition contents.
- `QueryAsOfTime` – Timestamp.

  The time as of when to read the partition contents. If not set, the most recent transaction commit time will be used. Cannot be specified along with `TransactionId`.

Response

- `Partitions` – An array of Partition objects.

  A list of requested partitions.
- `NextToken` – UTF-8 string.

A continuation token, if the returned list of partitions does not include the last one.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`
- `GlueEncryptionException`
- `InvalidStateException`
- `ResourceNotReadyException`
- `FederationSourceException`
- `FederationSourceRetryableException`

BatchGetPartition action (Python: batch_get_partition)

Retrieves partitions in a batch request.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog where the partitions in question reside. If none is supplied, the AWS account ID is used by default.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the catalog database where the partitions reside.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the partitions' table.
- `PartitionsToGet` – *Required:* An array of PartitionValueList objects, not more than 1000 structures.

A list of partition values identifying the partitions to retrieve.

Response

- `Partitions` – An array of Partition objects.

  A list of the requested partitions.
- `UnprocessedKeys` – An array of PartitionValueList objects, not more than 1000 structures.

  A list of the partition values in the request for which partitions were not returned.

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `OperationTimeoutException`
- `InternalServiceException`
- `GlueEncryptionException`
- `InvalidStateException`
- `FederationSourceException`
- `FederationSourceRetryableException`

BatchUpdatePartition action (Python: batch_update_partition)

Updates one or more partitions in a batch operation.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the catalog in which the partition is to be updated. Currently, this should be the AWS account ID.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The name of the metadata database in which the partition is to be updated.

- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the metadata table in which the partition is to be updated.
- `Entries` – *Required:* An array of BatchUpdatePartitionRequestEntry objects, not less than 1 or more than 100 structures.

  A list of up to 100 `BatchUpdatePartitionRequestEntry` objects to update.

Response

- `Errors` – An array of BatchUpdatePartitionFailureEntry objects.

  The errors encountered when trying to update the requested partitions. A list of `BatchUpdatePartitionFailureEntry` objects.

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `OperationTimeoutException`
- `InternalServiceException`
- `GlueEncryptionException`

GetColumnStatisticsForPartition action (Python: get_column_statistics_for_partition)

Retrieves partition statistics of columns.

The Identity and Access Management (IAM) permission required for this operation is `GetPartition`.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The ID of the Data Catalog where the partitions in question reside. If none is supplied, the AWS account ID is used by default.

- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the catalog database where the partitions reside.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the partitions' table.
- `PartitionValues` – *Required:* An array of UTF-8 strings.

  A list of partition values identifying the partition.
- `ColumnNames` – *Required:* An array of UTF-8 strings, not more than 100 strings.

  A list of the column names.

Response

- `ColumnStatisticsList` – An array of ColumnStatistics objects.

  List of ColumnStatistics that failed to be retrieved.
- `Errors` – An array of ColumnError objects.

  Error occurred during retrieving column statistics data.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `GlueEncryptionException`

UpdateColumnStatisticsForPartition action (Python: update_column_statistics_for_partition)

Creates or updates partition statistics of columns.

The Identity and Access Management (IAM) permission required for this operation is `UpdatePartition`.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog where the partitions in question reside. If none is supplied, the AWS account ID is used by default.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the catalog database where the partitions reside.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the partitions' table.
- `PartitionValues` – *Required:* An array of UTF-8 strings.

  A list of partition values identifying the partition.
- `ColumnStatisticsList` – *Required:* An array of ColumnStatistics objects, not more than 25 structures.

  A list of the column statistics.

Response

- `Errors` – An array of ColumnStatisticsError objects.

  Error occurred during updating column statistics data.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `GlueEncryptionException`

DeleteColumnStatisticsForPartition action (Python: delete_column_statistics_for_partition)

Delete the partition column statistics of a column.

The Identity and Access Management (IAM) permission required for this operation is `DeletePartition`.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog where the partitions in question reside. If none is supplied, the AWS account ID is used by default.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the catalog database where the partitions reside.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the partitions' table.
- `PartitionValues` – *Required:* An array of UTF-8 strings.

  A list of partition values identifying the partition.
- `ColumnName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Name of the column.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `GlueEncryptionException`

Connections API

The Connections API describes the data types and API related to working with connections in AWS Glue.

Topics

- Connection API
- Connection Types API
- Connection Metadata and Preview API

Connection API

The Connection API describes AWS Glue connection data types, and the API for creating, deleting, updating, and listing connections.

Data types

- Connection structure
- ConnectionInput structure
- TestConnectionInput structure
- PhysicalConnectionRequirements structure
- GetConnectionsFilter structure
- AuthenticationConfiguration structure

- AuthenticationConfigurationInput structure
- OAuth2Properties structure
- OAuth2PropertiesInput structure
- OAuth2ClientApplication structure
- AuthorizationCodeProperties structure
- BasicAuthenticationCredentials structure
- OAuth2Credentials structure

Connection structure

Defines a connection to a data source.

Fields

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the connection definition.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  The description of the connection.
- `ConnectionType` – UTF-8 string (valid values: `JDBC` | `SFTP` | `MONGODB` | `KAFKA` | `NETWORK` | `MARKETPLACE` | `CUSTOM` | `SALESFORCE` | `VIEW_VALIDATION_REDSHIFT` | `VIEW_VALIDATION_ATHENA` | `GOOGLEADS` | `GOOGLESHEETS` | `GOOGLEANALYTICS4` | `SERVICENOW` | `MARKETO` | `SAPODATA` | `ZENDESK` | `JIRACLOUD` | `NETSUITEERP` | `HUBSPOT` | `FACEBOOKADS` | `INSTAGRAMADS` | `ZOHOCRM` | `SALESFORCEPARDOT` | `SALESFORCEMARKETINGCLOUD` | `SLACK` | `STRIPE` | `INTERCOM` | `SNAPCHATADS`).

  The type of the connection. Currently, SFTP is not supported.
- `MatchCriteria` – An array of UTF-8 strings, not more than 10 strings.

  A list of criteria that can be used in selecting this connection.
- `ConnectionProperties` – A map array of key-value pairs, not more than 100 pairs.

  Each key is a UTF-8 string (valid values: `HOST` | `PORT` | `USERNAME="USER_NAME"` | `PASSWORD` | `ENCRYPTED_PASSWORD` | `JDBC_DRIVER_JAR_URI` | `JDBC_DRIVER_CLASS_NAME`

| `JDBC_ENGINE` | `JDBC_ENGINE_VERSION` | `CONFIG_FILES` | `INSTANCE_ID` | `JDBC_CONNECTION_URL` | `JDBC_ENFORCE_SSL` | `CUSTOM_JDBC_CERT` | `SKIP_CUSTOM_JDBC_CERT_VALIDATION` | `CUSTOM_JDBC_CERT_STRING` | `CONNECTION_URL` | `KAFKA_BOOTSTRAP_SERVERS` | `KAFKA_SSL_ENABLED` | `KAFKA_CUSTOM_CERT` | `KAFKA_SKIP_CUSTOM_CERT_VALIDATION` | `KAFKA_CLIENT_KEYSTORE` | `KAFKA_CLIENT_KEYSTORE_PASSWORD` | `KAFKA_CLIENT_KEY_PASSWORD` | `ENCRYPTED_KAFKA_CLIENT_KEYSTORE_PASSWORD` | `ENCRYPTED_KAFKA_CLIENT_KEY_PASSWORD` | `KAFKA_SASL_MECHANISM` | `KAFKA_SASL_PLAIN_USERNAME` | `KAFKA_SASL_PLAIN_PASSWORD` | `ENCRYPTED_KAFKA_SASL_PLAIN_PASSWORD` | `KAFKA_SASL_SCRAM_USERNAME` | `KAFKA_SASL_SCRAM_PASSWORD` | `KAFKA_SASL_SCRAM_SECRETS_ARN` | `ENCRYPTED_KAFKA_SASL_SCRAM_PASSWORD` | `KAFKA_SASL_GSSAPI_KEYTAB` | `KAFKA_SASL_GSSAPI_KRB5_CONF` | `KAFKA_SASL_GSSAPI_SERVICE` | `KAFKA_SASL_GSSAPI_PRINCIPAL` | `SECRET_ID` | `CONNECTOR_URL` | `CONNECTOR_TYPE` | `CONNECTOR_CLASS_NAME` | `ENDPOINT` | `ENDPOINT_TYPE` | `ROLE_ARN` | `REGION` | `WORKGROUP_NAME` | `CLUSTER_IDENTIFIER` | `DATABASE`).

Each value is a Value string, not less than 1 or more than 1024 bytes long.

These key-value pairs define parameters for the connection when using the version 1 Connection schema:

- `HOST` - The host URI: either the fully qualified domain name (FQDN) or the IPv4 address of the database host.
- `PORT` - The port number, between 1024 and 65535, of the port on which the database host is listening for database connections.
- `USER_NAME` - The name under which to log in to the database. The value string for `USER_NAME` is "`USERNAME`".
- `PASSWORD` - A password, if one is used, for the user name.
- `ENCRYPTED_PASSWORD` - When you enable connection password protection by setting `ConnectionPasswordEncryption` in the Data Catalog encryption settings, this field stores the encrypted password.
- `JDBC_DRIVER_JAR_URI` - The Amazon Simple Storage Service (Amazon S3) path of the JAR file that contains the JDBC driver to use.
- `JDBC_DRIVER_CLASS_NAME` - The class name of the JDBC driver to use.
- `JDBC_ENGINE` - The name of the JDBC engine to use.

- `JDBC_ENGINE_VERSION` - The version of the JDBC engine to use.
- `CONFIG_FILES` - (Reserved for future use.)
- `INSTANCE_ID` - The instance ID to use.
- `JDBC_CONNECTION_URL` - The URL for connecting to a JDBC data source.
- `JDBC_ENFORCE_SSL` - A case-insensitive Boolean string (true, false) specifying whether Secure Sockets Layer (SSL) with hostname matching is enforced for the JDBC connection on the client. The default is false.
- `CUSTOM_JDBC_CERT` - An Amazon S3 location specifying the customer's root certificate. AWS Glue uses this root certificate to validate the customer's certificate when connecting to the customer database. AWS Glue only handles X.509 certificates. The certificate provided must be DER-encoded and supplied in Base64 encoding PEM format.
- `SKIP_CUSTOM_JDBC_CERT_VALIDATION` - By default, this is `false`. AWS Glue validates the Signature algorithm and Subject Public Key Algorithm for the customer certificate. The only permitted algorithms for the Signature algorithm are SHA256withRSA, SHA384withRSA or SHA512withRSA. For the Subject Public Key Algorithm, the key length must be at least 2048. You can set the value of this property to `true` to skip AWS Glue's validation of the customer certificate.
- `CUSTOM_JDBC_CERT_STRING` - A custom JDBC certificate string which is used for domain match or distinguished name match to prevent a man-in-the-middle attack. In Oracle database, this is used as the `SSL_SERVER_CERT_DN`; in Microsoft SQL Server, this is used as the `hostNameInCertificate`.
- `CONNECTION_URL` - The URL for connecting to a general (non-JDBC) data source.
- `SECRET_ID` - The secret ID used for the secret manager of credentials.
- `CONNECTOR_URL` - The connector URL for a MARKETPLACE or CUSTOM connection.
- `CONNECTOR_TYPE` - The connector type for a MARKETPLACE or CUSTOM connection.
- `CONNECTOR_CLASS_NAME` - The connector class name for a MARKETPLACE or CUSTOM connection.
- `KAFKA_BOOTSTRAP_SERVERS` - A comma-separated list of host and port pairs that are the addresses of the Apache Kafka brokers in a Kafka cluster to which a Kafka client will connect to and bootstrap itself.
- `KAFKA_SSL_ENABLED` - Whether to enable or disable SSL on an Apache Kafka connection. Default value is "true".

- `KAFKA_CUSTOM_CERT` - The Amazon S3 URL for the private CA cert file (.pem format). The default is an empty string.
- `KAFKA_SKIP_CUSTOM_CERT_VALIDATION` - Whether to skip the validation of the CA cert file or not. AWS Glue validates for three algorithms: SHA256withRSA, SHA384withRSA and SHA512withRSA. Default value is "false".
- `KAFKA_CLIENT_KEYSTORE` - The Amazon S3 location of the client keystore file for Kafka client side authentication (Optional).
- `KAFKA_CLIENT_KEYSTORE_PASSWORD` - The password to access the provided keystore (Optional).
- `KAFKA_CLIENT_KEY_PASSWORD` - A keystore can consist of multiple keys, so this is the password to access the client key to be used with the Kafka server side key (Optional).
- `ENCRYPTED_KAFKA_CLIENT_KEYSTORE_PASSWORD` - The encrypted version of the Kafka client keystore password (if the user has the AWS Glue encrypt passwords setting selected).
- `ENCRYPTED_KAFKA_CLIENT_KEY_PASSWORD` - The encrypted version of the Kafka client key password (if the user has the AWS Glue encrypt passwords setting selected).
- `KAFKA_SASL_MECHANISM` - `"SCRAM-SHA-512"`, `"GSSAPI"`, `"AWS_MSK_IAM"`, or `"PLAIN"`. These are the supported SASL Mechanisms.
- `KAFKA_SASL_PLAIN_USERNAME` - A plaintext username used to authenticate with the "PLAIN" mechanism.
- `KAFKA_SASL_PLAIN_PASSWORD` - A plaintext password used to authenticate with the "PLAIN" mechanism.
- `ENCRYPTED_KAFKA_SASL_PLAIN_PASSWORD` - The encrypted version of the Kafka SASL PLAIN password (if the user has the AWS Glue encrypt passwords setting selected).
- `KAFKA_SASL_SCRAM_USERNAME` - A plaintext username used to authenticate with the "SCRAM-SHA-512" mechanism.
- `KAFKA_SASL_SCRAM_PASSWORD` - A plaintext password used to authenticate with the "SCRAM-SHA-512" mechanism.
- `ENCRYPTED_KAFKA_SASL_SCRAM_PASSWORD` - The encrypted version of the Kafka SASL SCRAM password (if the user has the AWS Glue encrypt passwords setting selected).
- `KAFKA_SASL_SCRAM_SECRETS_ARN` - The Amazon Resource Name of a secret in AWS Secrets Manager.

  - `KAFKA_SASL_GSSAPI_KEYTAB` - The S3 location of a Kerberos `keytab` file. A keytab stores long-term keys for one or more principals. For more information, see MIT Kerberos Documentation: Keytab.
  - `KAFKA_SASL_GSSAPI_KRB5_CONF` - The S3 location of a Kerberos `krb5.conf` file. A krb5.conf stores Kerberos configuration information, such as the location of the KDC server. For more information, see MIT Kerberos Documentation: krb5.conf.
  - `KAFKA_SASL_GSSAPI_SERVICE` - The Kerberos service name, as set with `sasl.kerberos.service.name` in your Kafka Configuration.
  - `KAFKA_SASL_GSSAPI_PRINCIPAL` - The name of the Kerberos princial used by AWS Glue. For more information, see Kafka Documentation: Configuring Kafka Brokers.
  - `ROLE_ARN` - The role to be used for running queries.
  - `REGION` - The AWS Region where queries will be run.
  - `WORKGROUP_NAME` - The name of an Amazon Redshift serverless workgroup or Amazon Athena workgroup in which queries will run.
  - `CLUSTER_IDENTIFIER` - The cluster identifier of an Amazon Redshift cluster in which queries will run.
  - `DATABASE` - The Amazon Redshift database that you are connecting to.
- `SparkProperties` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not less than 1 or more than 2048 bytes long.

  Connection properties specific to the Spark compute environment.
- `AthenaProperties` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not less than 1 or more than 2048 bytes long.

  Connection properties specific to the Athena compute environment.
- `PythonProperties` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not less than 1 or more than 2048 bytes long.

Connection properties specific to the Python compute environment.

- `PhysicalConnectionRequirements` – A PhysicalConnectionRequirements object.

  The physical connection requirements, such as virtual private cloud (VPC) and `SecurityGroup`, that are needed to make this connection successfully.
- `CreationTime` – Timestamp.

  The timestamp of the time that this connection definition was created.
- `LastUpdatedTime` – Timestamp.

  The timestamp of the last time the connection definition was updated.
- `LastUpdatedBy` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The user, group, or role that last updated this connection definition.
- `Status` – UTF-8 string (valid values: `READY` | `IN_PROGRESS` | `FAILED`).

  The status of the connection. Can be one of: `READY`, `IN_PROGRESS`, or `FAILED`.
- `StatusReason` – UTF-8 string, not less than 1 or more than 16384 bytes long.

  The reason for the connection status.
- `LastConnectionValidationTime` – Timestamp.

  A timestamp of the time this connection was last validated.
- `AuthenticationConfiguration` – An AuthenticationConfiguration object.

  The authentication properties of the connection.
- `ConnectionSchemaVersion` – Number (integer), not less than 1 or more than 2.

  The version of the connection schema for this connection. Version 2 supports properties for specific compute environments.
- `CompatibleComputeEnvironments` – An array of UTF-8 strings.

  A list of compute environments compatible with the connection.

ConnectionInput structure

A structure that is used to specify a connection to create or update.

Fields

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the connection.

- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  The description of the connection.

- `ConnectionType` – *Required:* UTF-8 string (valid values: `JDBC` | `SFTP` | `MONGODB` | `KAFKA` | `NETWORK` | `MARKETPLACE` | `CUSTOM` | `SALESFORCE` | `VIEW_VALIDATION_REDSHIFT` | `VIEW_VALIDATION_ATHENA` | `GOOGLEADS` | `GOOGLESHEETS` | `GOOGLEANALYTICS4` | `SERVICENOW` | `MARKETO` | `SAPODATA` | `ZENDESK` | `JIRACLOUD` | `NETSUITEERP` | `HUBSPOT` | `FACEBOOKADS` | `INSTAGRAMADS` | `ZOHOCRM` | `SALESFORCEPARDOT` | `SALESFORCEMARKETINGCLOUD` | `SLACK` | `STRIPE` | `INTERCOM` | `SNAPCHATADS`).

  The type of the connection. Currently, these types are supported:

  - `JDBC` - Designates a connection to a database through Java Database Connectivity (JDBC).

    `JDBC` Connections use the following ConnectionParameters.

    - Required: All of (`HOST`, `PORT`, `JDBC_ENGINE`) or `JDBC_CONNECTION_URL`.
    - Required: All of (`USERNAME`, `PASSWORD`) or `SECRET_ID`.
    - Optional: `JDBC_ENFORCE_SSL`, `CUSTOM_JDBC_CERT`, `CUSTOM_JDBC_CERT_STRING`, `SKIP_CUSTOM_JDBC_CERT_VALIDATION`. These parameters are used to configure SSL with JDBC.

  - `KAFKA` - Designates a connection to an Apache Kafka streaming platform.

    `KAFKA` Connections use the following ConnectionParameters.

    - Required: `KAFKA_BOOTSTRAP_SERVERS`.
    - Optional: `KAFKA_SSL_ENABLED`, `KAFKA_CUSTOM_CERT`, `KAFKA_SKIP_CUSTOM_CERT_VALIDATION`. These parameters are used to configure SSL with `KAFKA`.

  - Optional: `KAFKA_CLIENT_KEYSTORE`, `KAFKA_CLIENT_KEYSTORE_PASSWORD`, `KAFKA_CLIENT_KEY_PASSWORD`, `ENCRYPTED_KAFKA_CLIENT_KEYSTORE_PASSWORD`, `ENCRYPTED_KAFKA_CLIENT_KEY_PASSWORD`. These parameters are used to configure TLS client configuration with SSL in KAFKA.
  - Optional: `KAFKA_SASL_MECHANISM`. Can be specified as `SCRAM-SHA-512`, `GSSAPI`, or `AWS_MSK_IAM`.
  - Optional: `KAFKA_SASL_SCRAM_USERNAME`, `KAFKA_SASL_SCRAM_PASSWORD`, `ENCRYPTED_KAFKA_SASL_SCRAM_PASSWORD`. These parameters are used to configure SASL/SCRAM-SHA-512 authentication with KAFKA.
  - Optional: `KAFKA_SASL_GSSAPI_KEYTAB`, `KAFKA_SASL_GSSAPI_KRB5_CONF`, `KAFKA_SASL_GSSAPI_SERVICE`, `KAFKA_SASL_GSSAPI_PRINCIPAL`. These parameters are used to configure SASL/GSSAPI authentication with KAFKA.
- `MONGODB` - Designates a connection to a MongoDB document database.

  `MONGODB` Connections use the following ConnectionParameters.
  - Required: `CONNECTION_URL`.
  - Required: All of (`USERNAME`, `PASSWORD`) or `SECRET_ID`.
- `VIEW_VALIDATION_REDSHIFT` - Designates a connection used for view validation by Amazon Redshift.
- `VIEW_VALIDATION_ATHENA` - Designates a connection used for view validation by Amazon Athena.
- `NETWORK` - Designates a network connection to a data source within an Amazon Virtual Private Cloud environment (Amazon VPC).

  `NETWORK` Connections do not require ConnectionParameters. Instead, provide a PhysicalConnectionRequirements.
- `MARKETPLACE` - Uses configuration settings contained in a connector purchased from AWS Marketplace to read from and write to data stores that are not natively supported by AWS Glue.

  `MARKETPLACE` Connections use the following ConnectionParameters.
  - Required: `CONNECTOR_TYPE`, `CONNECTOR_URL`, `CONNECTOR_CLASS_NAME`, `CONNECTION_URL`.
  - Required for `JDBC` `CONNECTOR_TYPE` connections: All of (`USERNAME`, `PASSWORD`) or `SECRET_ID`.

- `CUSTOM` - Uses configuration settings contained in a custom connector to read from and write to data stores that are not natively supported by AWS Glue.

Additionally, a `ConnectionType` for the following SaaS connectors is supported:

- `FACEBOOKADS` - Designates a connection to Facebook Ads.
- `GOOGLEADS` - Designates a connection to Google Ads.
- `GOOGLESHEETS` - Designates a connection to Google Sheets.
- `GOOGLEANALYTICS4` - Designates a connection to Google Analytics 4.
- `HUBSPOT` - Designates a connection to HubSpot.
- `INSTAGRAMADS` - Designates a connection to Instagram Ads.
- `INTERCOM` - Designates a connection to Intercom.
- `JIRACLOUD` - Designates a connection to Jira Cloud.
- `MARKETO` - Designates a connection to Adobe Marketo Engage.
- `NETSUITEERP` - Designates a connection to Oracle NetSuite.
- `SALESFORCE` - Designates a connection to Salesforce using OAuth authentication.
- `SALESFORCEMARKETINGCLOUD` - Designates a connection to Salesforce Marketing Cloud.
- `SALESFORCEPARDOT` - Designates a connection to Salesforce Marketing Cloud Account Engagement (MCAE).
- `SAPODATA` - Designates a connection to SAP OData.
- `SERVICENOW` - Designates a connection to ServiceNow.
- `SLACK` - Designates a connection to Slack.
- `SNOWFLAKE` - Designates a connection to Snowflake.
- `SNAPCHATADS` - Designates a connection to Snapchat Ads.
- `STRIPE` - Designates a connection to Stripe.
- `ZENDESK` - Designates a connection to Zendesk.
- `ZOHOCRM` - Designates a connection to Zoho CRM.
- `ADOBEANALYTICS` - Designates a connection to Adobe Analytics.
- `LINKEDIN` - Designates a connection to LinkedIn.
- `MIXPANEL` - Designates a connection to Mixpanel.
- `ASANA` - Designates a connection to Asana.
- `SMARTSHEET` - Designates a connection to Smartsheet.

- `DATADOG` - Designates a connection to Datadog.
- `WOOCOMMERCE` - Designates a connection to WooCommerce.
- `PAYPAL` - Designates a connection to PayPal.
- `QUICKBOOKS` - Designates a connection to QuickBooks.
- `FACEBOOKPAGEINSIGHTS` - Designates a connection to Facebook Page Insights.
- `FRESHDESK` - Designates a connection to Freshdesk.
- `TWILIO` - Designates a connection to Twilio.
- `DOCUSIGNMONITOR` - Designates a connection to DocuSign Monitor.
- `FRESHSALES` - Designates a connection to Freshsales.
- `ZOOM` - Designates a connection to Zoom.
- `GOOGLESEARCHCONSOLE` - Designates a connection to Google Search Console.
- `SALESFORCECOMMERCECLOUD` - Designates a connection to Salesforce Commerce Cloud.
- `SAPCONCUR` - Designates a connection to SAP Concur.
- `DYNATRACE` - Designates a connection to Dynatrace.
- `MICROSOFTDYNAMIC365FINANCEANDOPS` - Designates a connection to Microsoft Dynamics 365 Finance and Operations.
- `MICROSOFTTEAMS` - Designates a connection to Microsoft Teams.
- `BLACKBAUDRAISEREDGENXT` - Designates a connection to Blackbaud Raiser's Edge NXT.
- `MAILCHIMP` - Designates a connection to Mailchimp.
- `GITLAB` - Designates a connection to GitLab.
- `PENDO` - Designates a connection to Pendo.
- `PRODUCTBOARD` - Designates a connection to Productboard.
- `CIRCLECI` - Designates a connection to CircleCI.
- `PIPEDIVE` - Designates a connection to Pipedrive.
- `SENDGRID` - Designates a connection to SendGrid.

For more information on the connection parameters needed for a particular connector, see the documentation for the connector in Adding an AWS Glue connectionin the AWS Glue User Guide.

SFTP is not supported.

For more information about how optional ConnectionProperties are used to configure features in AWS Glue, consult [AWS Glue connection properties](https://docs.aws.amazon.com/glue/latest/dg/connection-defining.html).

For more information about how optional ConnectionProperties are used to configure features in AWS Glue Studio, consult [Using connectors and connections](https://docs.aws.amazon.com/glue/latest/ug/connectors-chapter.html).

- `MatchCriteria` – An array of UTF-8 strings, not more than 10 strings.

  A list of criteria that can be used in selecting this connection.

- `ConnectionProperties` – *Required:* A map array of key-value pairs, not more than 100 pairs.

  Each key is a UTF-8 string (valid values: `HOST` | `PORT` | `USERNAME="USER_NAME"` | `PASSWORD` | `ENCRYPTED_PASSWORD` | `JDBC_DRIVER_JAR_URI` | `JDBC_DRIVER_CLASS_NAME` | `JDBC_ENGINE` | `JDBC_ENGINE_VERSION` | `CONFIG_FILES` | `INSTANCE_ID` | `JDBC_CONNECTION_URL` | `JDBC_ENFORCE_SSL` | `CUSTOM_JDBC_CERT` | `SKIP_CUSTOM_JDBC_CERT_VALIDATION` | `CUSTOM_JDBC_CERT_STRING` | `CONNECTION_URL` | `KAFKA_BOOTSTRAP_SERVERS` | `KAFKA_SSL_ENABLED` | `KAFKA_CUSTOM_CERT` | `KAFKA_SKIP_CUSTOM_CERT_VALIDATION` | `KAFKA_CLIENT_KEYSTORE` | `KAFKA_CLIENT_KEYSTORE_PASSWORD` | `KAFKA_CLIENT_KEY_PASSWORD` | `ENCRYPTED_KAFKA_CLIENT_KEYSTORE_PASSWORD` | `ENCRYPTED_KAFKA_CLIENT_KEY_PASSWORD` | `KAFKA_SASL_MECHANISM` | `KAFKA_SASL_PLAIN_USERNAME` | `KAFKA_SASL_PLAIN_PASSWORD` | `ENCRYPTED_KAFKA_SASL_PLAIN_PASSWORD` | `KAFKA_SASL_SCRAM_USERNAME` | `KAFKA_SASL_SCRAM_PASSWORD` | `KAFKA_SASL_SCRAM_SECRETS_ARN` | `ENCRYPTED_KAFKA_SASL_SCRAM_PASSWORD` | `KAFKA_SASL_GSSAPI_KEYTAB` | `KAFKA_SASL_GSSAPI_KRB5_CONF` | `KAFKA_SASL_GSSAPI_SERVICE` | `KAFKA_SASL_GSSAPI_PRINCIPAL` | `SECRET_ID` | `CONNECTOR_URL` | `CONNECTOR_TYPE` | `CONNECTOR_CLASS_NAME` | `ENDPOINT` | `ENDPOINT_TYPE` | `ROLE_ARN` | `REGION` | `WORKGROUP_NAME` | `CLUSTER_IDENTIFIER` | `DATABASE`).

  Each value is a Value string, not less than 1 or more than 1024 bytes long.

  These key-value pairs define parameters for the connection.

- `SparkProperties` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not less than 1 or more than 2048 bytes long.

  Connection properties specific to the Spark compute environment.

- `AthenaProperties` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not less than 1 or more than 2048 bytes long.

  Connection properties specific to the Athena compute environment.
- `PythonProperties` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not less than 1 or more than 2048 bytes long.

  Connection properties specific to the Python compute environment.
- `PhysicalConnectionRequirements` – A PhysicalConnectionRequirements object.

  The physical connection requirements, such as virtual private cloud (VPC) and `SecurityGroup`, that are needed to successfully make this connection.
- `AuthenticationConfiguration` – An AuthenticationConfigurationInput object.

  The authentication properties of the connection.
- `ValidateCredentials` – Boolean.

  A flag to validate the credentials during create connection. Default is true.
- `ValidateForComputeEnvironments` – An array of UTF-8 strings.

  The compute environments that the specified connection properties are validated against.

TestConnectionInput structure

A structure that is used to specify testing a connection to a service.

Fields

- `ConnectionType` – *Required:* UTF-8 string (valid values: `JDBC` | `SFTP` | `MONGODB` | `KAFKA` | `NETWORK` | `MARKETPLACE` | `CUSTOM` | `SALESFORCE` | `VIEW_VALIDATION_REDSHIFT` | `VIEW_VALIDATION_ATHENA` | `GOOGLEADS` | `GOOGLESHEETS` | `GOOGLEANALYTICS4`

| `SERVICENOW` | `MARKETO` | `SAPODATA` | `ZENDESK` | `JIRACLOUD` | `NETSUITEERP` | `HUBSPOT` | `FACEBOOKADS` | `INSTAGRAMADS` | `ZOHOCRM` | `SALESFORCEPARDOT` | `SALESFORCEMARKETINGCLOUD` | `SLACK` | `STRIPE` | `INTERCOM` | `SNAPCHATADS`).

The type of connection to test. This operation is only available for the `JDBC` or `SALESFORCE` connection types.

- `ConnectionProperties` – *Required:* A map array of key-value pairs, not more than 100 pairs.

  Each key is a UTF-8 string (valid values: `HOST` | `PORT` | `USERNAME="USER_NAME"` | `PASSWORD` | `ENCRYPTED_PASSWORD` | `JDBC_DRIVER_JAR_URI` | `JDBC_DRIVER_CLASS_NAME` | `JDBC_ENGINE` | `JDBC_ENGINE_VERSION` | `CONFIG_FILES` | `INSTANCE_ID` | `JDBC_CONNECTION_URL` | `JDBC_ENFORCE_SSL` | `CUSTOM_JDBC_CERT` | `SKIP_CUSTOM_JDBC_CERT_VALIDATION` | `CUSTOM_JDBC_CERT_STRING` | `CONNECTION_URL` | `KAFKA_BOOTSTRAP_SERVERS` | `KAFKA_SSL_ENABLED` | `KAFKA_CUSTOM_CERT` | `KAFKA_SKIP_CUSTOM_CERT_VALIDATION` | `KAFKA_CLIENT_KEYSTORE` | `KAFKA_CLIENT_KEYSTORE_PASSWORD` | `KAFKA_CLIENT_KEY_PASSWORD` | `ENCRYPTED_KAFKA_CLIENT_KEYSTORE_PASSWORD` | `ENCRYPTED_KAFKA_CLIENT_KEY_PASSWORD` | `KAFKA_SASL_MECHANISM` | `KAFKA_SASL_PLAIN_USERNAME` | `KAFKA_SASL_PLAIN_PASSWORD` | `ENCRYPTED_KAFKA_SASL_PLAIN_PASSWORD` | `KAFKA_SASL_SCRAM_USERNAME` | `KAFKA_SASL_SCRAM_PASSWORD` | `KAFKA_SASL_SCRAM_SECRETS_ARN` | `ENCRYPTED_KAFKA_SASL_SCRAM_PASSWORD` | `KAFKA_SASL_GSSAPI_KEYTAB` | `KAFKA_SASL_GSSAPI_KRB5_CONF` | `KAFKA_SASL_GSSAPI_SERVICE` | `KAFKA_SASL_GSSAPI_PRINCIPAL` | `SECRET_ID` | `CONNECTOR_URL` | `CONNECTOR_TYPE` | `CONNECTOR_CLASS_NAME` | `ENDPOINT` | `ENDPOINT_TYPE` | `ROLE_ARN` | `REGION` | `WORKGROUP_NAME` | `CLUSTER_IDENTIFIER` | `DATABASE`).

  Each value is a Value string, not less than 1 or more than 1024 bytes long.

  The key-value pairs that define parameters for the connection.

  JDBC connections use the following connection properties:

  - Required: All of (`HOST`, `PORT`, `JDBC_ENGINE`) or `JDBC_CONNECTION_URL`.
  - Required: All of (`USERNAME`, `PASSWORD`) or `SECRET_ID`.
  - Optional: `JDBC_ENFORCE_SSL`, `CUSTOM_JDBC_CERT`, `CUSTOM_JDBC_CERT_STRING`, `SKIP_CUSTOM_JDBC_CERT_VALIDATION`. These parameters are used to configure SSL with JDBC.

SALESFORCE connections require the `AuthenticationConfiguration` member to be configured.

- `AuthenticationConfiguration` – An AuthenticationConfigurationInput object.

  A structure containing the authentication configuration in the TestConnection request. Required for a connection to Salesforce using OAuth authentication.

PhysicalConnectionRequirements structure

The OAuth client app in GetConnection response.

Fields

- `SubnetId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The subnet ID used by the connection.
- `SecurityGroupIdList` – An array of UTF-8 strings, not more than 50 strings.

  The security group ID list used by the connection.
- `AvailabilityZone` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The connection's Availability Zone.

GetConnectionsFilter structure

Filters the connection definitions that are returned by the `GetConnections` API operation.

Fields

- `MatchCriteria` – An array of UTF-8 strings, not more than 10 strings.

  A criteria string that must match the criteria recorded in the connection definition for that connection definition to be returned.
- `ConnectionType` – UTF-8 string (valid values: `JDBC` | `SFTP` | `MONGODB` | `KAFKA` | `NETWORK` | `MARKETPLACE` | `CUSTOM` | `SALESFORCE` | `VIEW_VALIDATION_REDSHIFT` | `VIEW_VALIDATION_ATHENA` | `GOOGLEADS` | `GOOGLESHEETS` | `GOOGLEANALYTICS4`

| `SERVICENOW` | `MARKETO` | `SAPODATA` | `ZENDESK` | `JIRACLOUD` | `NETSUITEERP` | `HUBSPOT` | `FACEBOOKADS` | `INSTAGRAMADS` | `ZOHOCRM` | `SALESFORCEPARDOT` | `SALESFORCEMARKETINGCLOUD` | `SLACK` | `STRIPE` | `INTERCOM` | `SNAPCHATADS`).

The type of connections to return. Currently, SFTP is not supported.

- `ConnectionSchemaVersion` – Number (integer), not less than 1 or more than 2.

  Denotes if the connection was created with schema version 1 or 2.

AuthenticationConfiguration structure

A structure containing the authentication configuration.

Fields

- `AuthenticationType` – UTF-8 string (valid values: `BASIC` | `OAUTH2` | `CUSTOM` | `IAM`).

  A structure containing the authentication configuration.
- `SecretArn` – UTF-8 string, matching the Custom string pattern #36.

  The secret manager ARN to store credentials.
- `KmsKeyArn` – UTF-8 string, matching the Custom string pattern #29.

  The Amazon Resource Name (ARN) of the KMS key used to encrypt sensitive authentication information. This key is used to protect credentials and other sensitive data stored within the authentication configuration.
- `OAuth2Properties` – An OAuth2Properties object.

  The properties for OAuth2 authentication.

AuthenticationConfigurationInput structure

A structure containing the authentication configuration in the CreateConnection request.

Fields

- `AuthenticationType` – UTF-8 string (valid values: `BASIC` | `OAUTH2` | `CUSTOM` | `IAM`).

  A structure containing the authentication configuration in the CreateConnection request.
- `OAuth2Properties` – An OAuth2PropertiesInput object.

The properties for OAuth2 authentication in the CreateConnection request.

- `SecretArn` – UTF-8 string, matching the Custom string pattern #36.

  The secret manager ARN to store credentials in the CreateConnection request.
- `KmsKeyArn` – UTF-8 string, matching the Custom string pattern #29.

  The ARN of the KMS key used to encrypt the connection. Only taken an as input in the request and stored in the Secret Manager.
- `BasicAuthenticationCredentials` – A BasicAuthenticationCredentials object.

  The credentials used when the authentication type is basic authentication.
- `CustomAuthenticationCredentials` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not less than 1 or more than 2048 bytes long.

  The credentials used when the authentication type is custom authentication.

OAuth2Properties structure

A structure containing properties for OAuth2 authentication.

Fields

- `OAuth2GrantType` – UTF-8 string (valid values: `AUTHORIZATION_CODE` | `CLIENT_CREDENTIALS` | `JWT_BEARER`).

  The OAuth2 grant type. For example, `AUTHORIZATION_CODE`, `JWT_BEARER`, or `CLIENT_CREDENTIALS`.
- `OAuth2ClientApplication` – An OAuth2ClientApplication object.

  The client application type. For example, AWS_MANAGED or USER_MANAGED.
- `TokenUrl` – UTF-8 string, not more than 256 bytes long, matching the Custom string pattern #40.

  The URL of the provider's authentication server, to exchange an authorization code for an access token.

- `TokenUrlParametersMap` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not less than 1 or more than 512 bytes long.

  A map of parameters that are added to the token `GET` request.

OAuth2PropertiesInput structure

A structure containing properties for OAuth2 in the CreateConnection request.

Fields

- `OAuth2GrantType` – UTF-8 string (valid values: `AUTHORIZATION_CODE` | `CLIENT_CREDENTIALS` | `JWT_BEARER`).

  The OAuth2 grant type in the CreateConnection request. For example, `AUTHORIZATION_CODE`, `JWT_BEARER`, or `CLIENT_CREDENTIALS`.
- `OAuth2ClientApplication` – An OAuth2ClientApplication object.

  The client application type in the CreateConnection request. For example, `AWS_MANAGED` or `USER_MANAGED`.
- `TokenUrl` – UTF-8 string, not more than 256 bytes long, matching the Custom string pattern #40.

  The URL of the provider's authentication server, to exchange an authorization code for an access token.
- `TokenUrlParametersMap` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not less than 1 or more than 512 bytes long.

  A map of parameters that are added to the token `GET` request.
- `AuthorizationCodeProperties` – An AuthorizationCodeProperties object.

  The set of properties required for the the OAuth2 `AUTHORIZATION_CODE` grant type.
- `OAuth2Credentials` – An OAuth2Credentials object.

The credentials used when the authentication type is OAuth2 authentication.

OAuth2ClientApplication structure

The OAuth2 client app used for the connection.

Fields

- `UserManagedClientApplicationClientId` – UTF-8 string, not more than 2048 bytes long, matching the Custom string pattern #37.

  The client application clientID if the ClientAppType is `USER_MANAGED`.
- `AWSManagedClientApplicationReference` – UTF-8 string, not more than 2048 bytes long, matching the Custom string pattern #37.

  The reference to the SaaS-side client app that is AWS managed.

AuthorizationCodeProperties structure

The set of properties required for the the OAuth2 `AUTHORIZATION_CODE` grant type workflow.

Fields

- `AuthorizationCode` – UTF-8 string, not less than 1 or more than 4096 bytes long, matching the Custom string pattern #37.

  An authorization code to be used in the third leg of the `AUTHORIZATION_CODE` grant workflow. This is a single-use code which becomes invalid once exchanged for an access token, thus it is acceptable to have this value as a request parameter.
- `RedirectUri` – UTF-8 string, not more than 512 bytes long, matching the Custom string pattern #41.

  The redirect URI where the user gets redirected to by authorization server when issuing an authorization code. The URI is subsequently used when the authorization code is exchanged for an access token.

BasicAuthenticationCredentials structure

For supplying basic auth credentials when not providing a `SecretArn` value.

Fields

- `Username` – UTF-8 string, not more than 512 bytes long, matching the Custom string pattern #37.

  The username to connect to the data source.
- `Password` – UTF-8 string, not more than 512 bytes long, matching the Custom string pattern #33.

  The password to connect to the data source.

OAuth2Credentials structure

The credentials used when the authentication type is OAuth2 authentication.

Fields

- `UserManagedClientApplicationClientSecret` – UTF-8 string, not more than 512 bytes long, matching the Custom string pattern #38.

  The client application client secret if the client application is user managed.
- `AccessToken` – UTF-8 string, not more than 4096 bytes long, matching the Custom string pattern #38.

  The access token used when the authentication type is OAuth2.
- `RefreshToken` – UTF-8 string, not more than 4096 bytes long, matching the Custom string pattern #38.

  The refresh token used when the authentication type is OAuth2.
- `JwtToken` – UTF-8 string, not more than 8000 bytes long, matching the Custom string pattern #39.

  The JSON Web Token (JWT) used when the authentication type is OAuth2.

Operations

- CreateConnection action (Python: create_connection)
- DeleteConnection action (Python: delete_connection)

- GetConnection action (Python: get_connection)
- GetConnections action (Python: get_connections)
- UpdateConnection action (Python: update_connection)
- TestConnection action (Python: test_connection)
- BatchDeleteConnection action (Python: batch_delete_connection)

CreateConnection action (Python: create_connection)

Creates a connection definition in the Data Catalog.

Connections used for creating federated resources require the IAM `glue:PassConnection` permission.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog in which to create the connection. If none is provided, the AWS account ID is used by default.
- `ConnectionInput` – *Required:* A ConnectionInput object.

  A `ConnectionInput` object defining the connection to create.
- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  The tags you assign to the connection.

Response

- `CreateConnectionStatus` – UTF-8 string (valid values: `READY` | `IN_PROGRESS` | `FAILED`).

  The status of the connection creation request. The request can take some time for certain authentication types, for example when creating an OAuth connection with token exchange over VPC.

Errors

- `AlreadyExistsException`
- `InvalidInputException`
- `OperationTimeoutException`
- `ResourceNumberLimitExceededException`
- `GlueEncryptionException`

DeleteConnection action (Python: delete_connection)

Deletes a connection from the Data Catalog.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog in which the connection resides. If none is provided, the AWS account ID is used by default.
- `ConnectionName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the connection to delete.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `OperationTimeoutException`

GetConnection action (Python: get_connection)

Retrieves a connection definition from the Data Catalog.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog in which the connection resides. If none is provided, the AWS account ID is used by default.

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the connection definition to retrieve.

- `HidePassword` – Boolean.

  Allows you to retrieve the connection metadata without returning the password. For instance, the AWS Glue console uses this flag to retrieve the connection, and does not display the password. Set this parameter when the caller might not have permission to use the AWS KMS key to decrypt the password, but it does have permission to access the rest of the connection properties.

- `ApplyOverrideForComputeEnvironment` – UTF-8 string (valid values: `SPARK` | `ATHENA` | `PYTHON`).

  For connections that may be used in multiple services, specifies returning properties for the specified compute environment.

Response

- `Connection` – A Connection object.

  The requested connection definition.

Errors

- `EntityNotFoundException`
- `OperationTimeoutException`
- `InvalidInputException`
- `GlueEncryptionException`

GetConnections action (Python: get_connections)

Retrieves a list of connection definitions from the Data Catalog.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog in which the connections reside. If none is provided, the AWS account ID is used by default.
- `Filter` – A GetConnectionsFilter object.

  A filter that controls which connections are returned.
- `HidePassword` – Boolean.

  Allows you to retrieve the connection metadata without returning the password. For instance, the AWS Glue console uses this flag to retrieve the connection, and does not display the password. Set this parameter when the caller might not have permission to use the AWS KMS key to decrypt the password, but it does have permission to access the rest of the connection properties.
- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation call.
- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum number of connections to return in one response.

Response

- `ConnectionList` – An array of Connection objects.

  A list of requested connection definitions.
- `NextToken` – UTF-8 string.

  A continuation token, if the list of connections returned does not include the last of the filtered connections.

Errors

- `EntityNotFoundException`
- `OperationTimeoutException`
- `InvalidInputException`
- `GlueEncryptionException`

UpdateConnection action (Python: update_connection)

Updates a connection definition in the Data Catalog.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog in which the connection resides. If none is provided, the AWS account ID is used by default.
- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the connection definition to update.
- `ConnectionInput` – *Required:* A ConnectionInput object.

  A `ConnectionInput` object that redefines the connection in question.

Response

- *No Response parameters.*

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `OperationTimeoutException`
- `InvalidInputException`
- `GlueEncryptionException`

TestConnection action (Python: test_connection)

Tests a connection to a service to validate the service credentials that you provide.

You can either provide an existing connection name or a `TestConnectionInput` for testing a non-existing connection input. Providing both at the same time will cause an error.

If the action is successful, the service sends back an HTTP 200 response.

Request

- `ConnectionName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Optional. The name of the connection to test. If only name is provided, the operation will get the connection and use that for testing.
- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The catalog ID where the connection resides.
- `TestConnectionInput` – A TestConnectionInput object.

  A structure that is used to specify testing a connection to a service.

Response

- *No Response parameters.*

Errors

- `InvalidInputException`
- `OperationTimeoutException`
- `ResourceNumberLimitExceededException`
- `GlueEncryptionException`
- `FederationSourceException`
- `AccessDeniedException`
- `EntityNotFoundException`
- `ConflictException`

- `InternalServiceException`

BatchDeleteConnection action (Python: batch_delete_connection)

Deletes a list of connection definitions from the Data Catalog.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog in which the connections reside. If none is provided, the AWS account ID is used by default.
- `ConnectionNameList` – *Required:* An array of UTF-8 strings, not more than 25 strings.

  A list of names of the connections to delete.

Response

- `Succeeded` – An array of UTF-8 strings.

  A list of names of the connection definitions that were successfully deleted.
- `Errors` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a An ErrorDetail object.

  A map of the names of connections that were not successfully deleted to error details.

Errors

- `InternalServiceException`
- `OperationTimeoutException`

Connection Types API

The Connection Type API describes AWS Glue APIs related to describing connection types.

Connection management APIs

- DescribeConnectionType action (Python: describe_connection_type)
- ListConnectionTypes action (Python: list_connection_types)
- ConnectionTypeBrief structure
- ConnectionTypeVariant structure

DescribeConnectionType action (Python: describe_connection_type)

The `DescribeConnectionType` API provides full details of the supported options for a given connection type in AWS Glue.

Request

- `ConnectionType` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the connection type to be described.

Response

- `ConnectionType` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the connection type.
- `Description` – UTF-8 string, not more than 1024 bytes long.

  A description of the connection type.
- `Capabilities` – A Capabilities object.

  The supported authentication types, data interface types (compute environments), and data operations of the connector.
- `ConnectionProperties` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a A Property object.

  Connection properties which are common across compute environments.

- `ConnectionOptions` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a A Property object.

  Returns properties that can be set when creating a connection in the `ConnectionInput.ConnectionProperties`. `ConnectionOptions` defines parameters that can be set in a Spark ETL script in the connection options map passed to a dataframe.
- `AuthenticationConfiguration` – An AuthConfiguration object.

  The type of authentication used for the connection.
- `ComputeEnvironmentConfigurations` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a A ComputeEnvironmentConfiguration object.

  The compute environments that are supported by the connection.
- `PhysicalConnectionRequirements` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a A Property object.

  Physical requirements for a connection, such as VPC, Subnet and Security Group specifications.
- `AthenaConnectionProperties` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a A Property object.

  Connection properties specific to the Athena compute environment.
- `PythonConnectionProperties` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a A Property object.

  Connection properties specific to the Python compute environment.
- `SparkConnectionProperties` – A map array of key-value pairs.

Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

Each value is a A Property object.

Connection properties specific to the Spark compute environment.

Errors

- `ValidationException`
- `InvalidInputException`
- `InternalServiceException`

ListConnectionTypes action (Python: list_connection_types)

The `ListConnectionTypes` API provides a discovery mechanism to learn available connection types in AWS Glue. The response contains a list of connection types with high-level details of what is supported for each connection type. The connection types listed are the set of supported options for the `ConnectionType` value in the `CreateConnection` API.

Request

- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum number of results to return.

- `NextToken` – UTF-8 string, not less than 1 or more than 2048 bytes long, matching the Custom string pattern #11.

  A continuation token, if this is a continuation call.

Response

- `ConnectionTypes` – An array of ConnectionTypeBrief objects.

  A list of `ConnectionTypeBrief` objects containing brief information about the supported connection types.

- `NextToken` – UTF-8 string, not less than 1 or more than 2048 bytes long, matching the Custom string pattern #11.

A continuation token, if the current list segment is not the last.

Errors

- `InternalServiceException`

ConnectionTypeBrief structure

Brief information about a supported connection type returned by the `ListConnectionTypes` API.

Fields

- `ConnectionType` – UTF-8 string (valid values: `JDBC` | `SFTP` | `MONGODB` | `KAFKA` | `NETWORK` | `MARKETPLACE` | `CUSTOM` | `SALESFORCE` | `VIEW_VALIDATION_REDSHIFT` | `VIEW_VALIDATION_ATHENA` | `GOOGLEADS` | `GOOGLESHEETS` | `GOOGLEANALYTICS4` | `SERVICENOW` | `MARKETO` | `SAPODATA` | `ZENDESK` | `JIRACLOUD` | `NETSUITEERP` | `HUBSPOT` | `FACEBOOKADS` | `INSTAGRAMADS` | `ZOHOCRM` | `SALESFORCEPARDOT` | `SALESFORCEMARKETINGCLOUD` | `SLACK` | `STRIPE` | `INTERCOM` | `SNAPCHATADS`).

  The name of the connection type.
- `DisplayName` – UTF-8 string, not less than 1 or more than 128 bytes long.

  The human-readable name for the connection type that is displayed in the AWS Glue console.
- `Vendor` – UTF-8 string, not less than 1 or more than 128 bytes long.

  The name of the vendor or provider that created or maintains this connection type.
- `Description` – UTF-8 string, not more than 1024 bytes long.

  A description of the connection type.
- `Categories` – An array of UTF-8 strings.

  A list of categories that this connection type belongs to. Categories help users filter and find appropriate connection types based on their use cases.
- `Capabilities` – A Capabilities object.

  The supported authentication types, data interface types (compute environments), and data operations of the connector.

- `LogoUrl` – UTF-8 string.

  The URL of the logo associated with a connection type.
- `ConnectionTypeVariants` – An array of [ConnectionTypeVariant](#) objects.

  A list of variants available for this connection type. Different variants may provide specialized configurations for specific use cases or implementations of the same general connection type.

ConnectionTypeVariant structure

Represents a variant of a connection type in AWS Glue Data Catalog. Connection type variants provide specific configurations and behaviors for different implementations of the same general connection type.

Fields

- `ConnectionTypeVariantName` – UTF-8 string, not less than 1 or more than 128 bytes long.

  The unique identifier for the connection type variant. This name is used internally to identify the specific variant of a connection type.
- `DisplayName` – UTF-8 string, not less than 1 or more than 128 bytes long.

  The human-readable name for the connection type variant that is displayed in the AWS Glue console.
- `Description` – UTF-8 string, not more than 1024 bytes long.

  A detailed description of the connection type variant, including its purpose, use cases, and any specific configuration requirements.
- `LogoUrl` – UTF-8 string.

  The URL of the logo associated with a connection type variant.

datatypes

- [Validation structure](#)
- [AuthConfiguration structure](#)
- [Capabilities structure](#)
- [Property structure](#)

- AllowedValue structure
- ComputeEnvironmentConfiguration structure

Validation structure

Defines how a validation is performed on a connection property.

Fields

- `ValidationType` – *Required:* UTF-8 string (valid values: `REGEX` | `RANGE`).

  The type of validation to be performed, such as `REGEX`.
- `Patterns` – An array of UTF-8 strings.

  A list of patterns that apply to the validation.
- `Description` – *Required:* UTF-8 string, not less than 1 or more than 1024 bytes long.

  A description for the validation.
- `MaxLength` – Number (integer).

  A maximum length for a string connection property.
- `Maximum` – Number (integer).

  A maximum value when specifying a `RANGE` type of validation.
- `Minimum` – Number (integer).

  A minimum value when specifying a `RANGE` type of validation.

AuthConfiguration structure

The authentication configuration for a connection returned by the `DescribeConnectionType` API.

Fields

- `AuthenticationType` – *Required:* A Property object.

  The type of authentication for a connection.
- `SecretArn` – A Property object.

The Amazon Resource Name (ARN) for the Secrets Manager.

- `OAuth2Properties` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a A Property object.

  A map of key-value pairs for the OAuth2 properties. Each value is a a `Property` object.
- `BasicAuthenticationProperties` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a A Property object.

  A map of key-value pairs for the OAuth2 properties. Each value is a a `Property` object.
- `CustomAuthenticationProperties` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a A Property object.

  A map of key-value pairs for the custom authentication properties. Each value is a a `Property` object.

Capabilities structure

Specifies the supported authentication types returned by the `DescribeConnectionType` API.

Fields

- `SupportedAuthenticationTypes` – *Required:* An array of UTF-8 strings.

  A list of supported authentication types.
- `SupportedDataOperations` – *Required:* An array of UTF-8 strings.

  A list of supported data operations.
- `SupportedComputeEnvironments` – *Required:* An array of UTF-8 strings.

  A list of supported compute environments.

Property structure

An object that defines a connection type for a compute environment.

Fields

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The name of the property.
- `Description` – *Required:* UTF-8 string, not more than 1024 bytes long.

  A description of the property.
- `Required` – *Required:* Boolean.

  Indicates whether the property is required.
- `DefaultValue` – UTF-8 string.

  The default value for the property.
- `PropertyTypes` – *Required:* An array of UTF-8 strings.

  Describes the type of property.
- `AllowedValues` – An array of AllowedValue objects.

  A list of `AllowedValue` objects representing the values allowed for the property.
- `DataOperationScopes` – An array of UTF-8 strings.

  Indicates which data operations are applicable to the property.

AllowedValue structure

An object representing a value allowed for a property.

Fields

- `Description` – UTF-8 string, not more than 1024 bytes long.

  A description of the allowed value.
- `Value` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The value allowed for the property.

ComputeEnvironmentConfiguration structure

An object containing configuration for a compute environment (such as Spark, Python or Athena) returned by the `DescribeConnectionType` API.

Fields

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  A name for the compute environment configuration.
- `Description` – *Required:* UTF-8 string, not more than 1024 bytes long.

  A description of the compute environment.
- `ComputeEnvironment` – *Required:* UTF-8 string (valid values: `SPARK` | `ATHENA` | `PYTHON`).

  The type of compute environment.
- `SupportedAuthenticationTypes` – *Required:* An array of UTF-8 strings.

  The supported authentication types for the compute environment.
- `ConnectionOptions` – *Required:* A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a A Property object.

  The parameters used as connection options for the compute environment.
- `ConnectionPropertyNameOverrides` – *Required:* A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not less than 1 or more than 128 bytes long.

  The connection property name overrides for the compute environment.
- `ConnectionOptionNameOverrides` – *Required:* A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not less than 1 or more than 128 bytes long.

  The connection option name overrides for the compute environment.
- `ConnectionPropertiesRequiredOverrides` – *Required:* An array of UTF-8 strings.

The connection properties that are required as overrides for the compute environment.

- `PhysicalConnectionPropertiesRequired` – Boolean.

  Indicates whether `PhysicalConnectionProperties` are required for the compute environment.

Connection Metadata and Preview API

The following connection APIs describe operations for describing connection metadata.

Data types

- Entity structure
- Field structure

Entity structure

An entity supported by a given `ConnectionType`.

Fields

- `EntityName` – UTF-8 string.

  The name of the entity.
- `Label` – UTF-8 string.

  Label used for the entity.
- `IsParentEntity` – Boolean.

  A Boolean value which helps to determine whether there are sub objects that can be listed.
- `Description` – UTF-8 string.

  A description of the entity.
- `Category` – UTF-8 string.

  The type of entities that are present in the response. This value depends on the source connection. For example this is `SObjects` for Salesforce and `databases` or `schemas` or `tables` for sources like Amazon Redshift.

- `CustomProperties` – A map array of key-value pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

  An optional map of keys which may be returned for an entity by a connector.

Field structure

The `Field` object has information about the different properties associated with a field in the connector.

Fields

- `FieldName` – UTF-8 string.

  A unique identifier for the field.
- `Label` – UTF-8 string.

  A readable label used for the field.
- `Description` – UTF-8 string.

  A description of the field.
- `FieldType` – UTF-8 string (valid values: `INT` | `SMALLINT` | `BIGINT` | `FLOAT` | `LONG` | `DATE` | `BOOLEAN` | `MAP` | `ARRAY` | `STRING` | `TIMESTAMP` | `DECIMAL` | `BYTE` | `SHORT` | `DOUBLE` | `STRUCT`).

  The type of data in the field.
- `IsPrimaryKey` – Boolean.

  Indicates whether this field can used as a primary key for the given entity.
- `IsNullable` – Boolean.

  Indicates whether this field can be nullable or not.
- `IsRetrievable` – Boolean.

  Indicates whether this field can be added in Select clause of SQL query or whether it is retrievable or not.
- `IsFilterable` – Boolean.

Indicates whether this field can used in a filter clause (`WHERE` clause) of a SQL statement when querying data.

- `IsPartitionable` – Boolean.

  Indicates whether a given field can be used in partitioning the query made to SaaS.

- `IsCreateable` – Boolean.

  Indicates whether this field can be created as part of a destination write.

- `IsUpdateable` – Boolean.

  Indicates whether this field can be updated as part of a destination write.

- `IsUpsertable` – Boolean.

  Indicates whether this field can be upserted as part of a destination write.

- `IsDefaultOnCreate` – Boolean.

  Indicates whether this field is populated automatically when the object is created, such as a created at timestamp.

- `SupportedValues` – An array of UTF-8 strings.

  A list of supported values for the field.

- `SupportedFilterOperators` – An array of UTF-8 strings.

  Indicates the support filter operators for this field.

- `ParentField` – UTF-8 string.

  A parent field name for a nested field.

- `NativeDataType` – UTF-8 string.

  The data type returned by the SaaS API, such as "picklist" or "textarea" from Salesforce.

- `CustomProperties` – A map array of key-value pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

  Optional map of keys which may be returned.

Operations

- ListEntities action (Python: list_entities)
- DescribeEntity action (Python: describe_entity)
- GetEntityRecords action (Python: get_entity_records)

ListEntities action (Python: list_entities)

Returns the available entities supported by the connection type.

Request

- `ConnectionName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A name for the connection that has required credentials to query any connection type.
- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The catalog ID of the catalog that contains the connection. This can be null, By default, the AWS Account ID is the catalog ID.
- `ParentEntityName` – UTF-8 string.

  Name of the parent entity for which you want to list the children. This parameter takes a fully-qualified path of the entity in order to list the child entities.
- `NextToken` – UTF-8 string, not less than 1 or more than 2048 bytes long, matching the Custom string pattern #11.

  A continuation token, included if this is a continuation call.
- `DataStoreApiVersion` – UTF-8 string, not less than 1 or more than 256 bytes long, matching the Custom string pattern #23.

  The API version of the SaaS connector.

Response

- `Entities` – An array of Entity objects.

A list of `Entity` objects.

- `NextToken` – UTF-8 string, not less than 1 or more than 2048 bytes long, matching the Custom string pattern #11.

  A continuation token, present if the current segment is not the last.

Errors

- `EntityNotFoundException`
- `OperationTimeoutException`
- `InvalidInputException`
- `GlueEncryptionException`
- `ValidationException`
- `FederationSourceException`
- `AccessDeniedException`

DescribeEntity action (Python: describe_entity)

Provides details regarding the entity used with the connection type, with a description of the data model for each field in the selected entity.

The response includes all the fields which make up the entity.

Request

- `ConnectionName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the connection that contains the connection type credentials.
- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The catalog ID of the catalog that contains the connection. This can be null, By default, the AWS Account ID is the catalog ID.
- `EntityName` – *Required:* UTF-8 string.

  The name of the entity that you want to describe from the connection type.

- `NextToken` – UTF-8 string, not less than 1 or more than 2048 bytes long, matching the Custom string pattern #11.

  A continuation token, included if this is a continuation call.
- `DataStoreApiVersion` – UTF-8 string, not less than 1 or more than 256 bytes long, matching the Custom string pattern #23.

  The version of the API used for the data store.

Response

- `Fields` – An array of Field objects.

  Describes the fields for that connector entity. This is the list of `Field` objects. `Field` is very similar to column in a database. The `Field` object has information about different properties associated with fields in the connector.
- `NextToken` – UTF-8 string, not less than 1 or more than 2048 bytes long, matching the Custom string pattern #11.

  A continuation token, present if the current segment is not the last.

Errors

- `EntityNotFoundException`
- `OperationTimeoutException`
- `InvalidInputException`
- `GlueEncryptionException`
- `ValidationException`
- `FederationSourceException`
- `AccessDeniedException`

GetEntityRecords action (Python: get_entity_records)

This API is used to query preview data from a given connection type or from a native Amazon S3 based AWS Glue Data Catalog.

Returns records as an array of JSON blobs. Each record is formatted using Jackson JsonNode based on the field type defined by the `DescribeEntity` API.

Spark connectors generate schemas according to the same data type mapping as in the `DescribeEntity` API. Spark connectors convert data to the appropriate data types matching the schema when returning rows.

Request

- `ConnectionName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the connection that contains the connection type credentials.
- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The catalog ID of the catalog that contains the connection. This can be null, By default, the AWS Account ID is the catalog ID.
- `EntityName` – *Required:* UTF-8 string.

  Name of the entity that we want to query the preview data from the given connection type.
- `NextToken` – UTF-8 string, not less than 1 or more than 2048 bytes long, matching the Custom string pattern #11.

  A continuation token, included if this is a continuation call.
- `DataStoreApiVersion` – UTF-8 string, not less than 1 or more than 256 bytes long, matching the Custom string pattern #23.

  The API version of the SaaS connector.
- `ConnectionOptions` – A map array of key-value pairs, not more than 100 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 256 bytes long, matching the Custom string pattern #18.

  Each value is a UTF-8 string, not less than 1 or more than 256 bytes long, matching the Custom string pattern #17.

  Connector options that are required to query the data.
- `FilterPredicate` – UTF-8 string, not less than 1 or more than 100000 bytes long.

A filter predicate that you can apply in the query request.

- `Limit` – *Required:* Number (long), not less than 1 or more than 1000.

  Limits the number of records fetched with the request.
- `OrderBy` – UTF-8 string.

  A parameter that orders the response preview data.
- `SelectedFields` – An array of UTF-8 strings, not less than 1 or more than 1000 strings.

  List of fields that we want to fetch as part of preview data.

Response

- `Records` – An array of a structures.

  A list of the requested objects.
- `NextToken` – UTF-8 string, not less than 1 or more than 2048 bytes long, matching the Custom string pattern #11.

  A continuation token, present if the current segment is not the last.

Errors

- `EntityNotFoundException`
- `OperationTimeoutException`
- `InvalidInputException`
- `GlueEncryptionException`
- `ValidationException`
- `FederationSourceException`
- `AccessDeniedException`

User-defined Function API

The User-defined Function API describes AWS Glue data types and operations used in working with functions.

Data types

- UserDefinedFunction structure
- UserDefinedFunctionInput structure

UserDefinedFunction structure

Represents the equivalent of a Hive user-defined function (UDF) definition.

Fields

- `FunctionName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the function.
- `DatabaseName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the catalog database that contains the function.
- `ClassName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Java class that contains the function code.
- `OwnerName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The owner of the function.
- `OwnerType` – UTF-8 string (valid values: `USER` | `ROLE` | `GROUP`).

  The owner type.
- `CreateTime` – Timestamp.

  The time at which the function was created.
- `ResourceUris` – An array of ResourceUri objects, not more than 1000 structures.

  The resource URIs for the function.
- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The ID of the Data Catalog in which the function resides.

UserDefinedFunctionInput structure

A structure used to create or update a user-defined function.

Fields

- `FunctionName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the function.
- `ClassName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Java class that contains the function code.
- `OwnerName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The owner of the function.
- `OwnerType` – UTF-8 string (valid values: `USER` | `ROLE` | `GROUP`).

  The owner type.
- `ResourceUris` – An array of ResourceUri objects, not more than 1000 structures.

  The resource URIs for the function.

Operations

- CreateUserDefinedFunction action (Python: create_user_defined_function)
- UpdateUserDefinedFunction action (Python: update_user_defined_function)
- DeleteUserDefinedFunction action (Python: delete_user_defined_function)
- GetUserDefinedFunction action (Python: get_user_defined_function)
- GetUserDefinedFunctions action (Python: get_user_defined_functions)

CreateUserDefinedFunction action (Python: create_user_defined_function)

Creates a new function definition in the Data Catalog.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog in which to create the function. If none is provided, the AWS account ID is used by default.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the catalog database in which to create the function.
- `FunctionInput` – *Required:* An UserDefinedFunctionInput object.

  A `FunctionInput` object that defines the function to create in the Data Catalog.

Response

- *No Response parameters.*

Errors

- `AlreadyExistsException`
- `InvalidInputException`
- `InternalServiceException`
- `EntityNotFoundException`
- `OperationTimeoutException`
- `ResourceNumberLimitExceededException`
- `GlueEncryptionException`

UpdateUserDefinedFunction action (Python: update_user_defined_function)

Updates an existing function definition in the Data Catalog.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog where the function to be updated is located. If none is provided, the AWS account ID is used by default.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the catalog database where the function to be updated is located.
- `FunctionName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the function.
- `FunctionInput` – *Required:* An UserDefinedFunctionInput object.

  A `FunctionInput` object that redefines the function in the Data Catalog.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `GlueEncryptionException`

DeleteUserDefinedFunction action (Python: delete_user_defined_function)

Deletes an existing function definition from the Data Catalog.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog where the function to be deleted is located. If none is supplied, the AWS account ID is used by default.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the catalog database where the function is located.
- `FunctionName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the function definition to be deleted.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`

GetUserDefinedFunction action (Python: get_user_defined_function)

Retrieves a specified function definition from the Data Catalog.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog where the function to be retrieved is located. If none is provided, the AWS account ID is used by default.

- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the catalog database where the function is located.
- `FunctionName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the function.

Response

- `UserDefinedFunction` – An UserDefinedFunction object.

  The requested function definition.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `GlueEncryptionException`

GetUserDefinedFunctions action (Python: get_user_defined_functions)

Retrieves multiple function definitions from the Data Catalog.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog where the functions to be retrieved are located. If none is provided, the AWS account ID is used by default.
- `DatabaseName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The name of the catalog database where the functions are located. If none is provided, functions from all the databases across the catalog will be returned.

- `Pattern` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  An optional function-name pattern string that filters the function definitions returned.
- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation call.
- `MaxResults` – Number (integer), not less than 1 or more than 100.

  The maximum number of functions to return in one response.

Response

- `UserDefinedFunctions` – An array of UserDefinedFunction objects.

  A list of requested function definitions.
- `NextToken` – UTF-8 string.

  A continuation token, if the list of functions returned does not include the last requested function.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`
- `GlueEncryptionException`

Importing an Athena catalog to AWS Glue

The Migration API describes AWS Glue data types and operations having to do with migrating an Athena Data catalog to AWS Glue.

Data types

- CatalogImportStatus structure

CatalogImportStatus structure

A structure containing migration status information.

Fields

- `ImportCompleted` – Boolean.

  `True` if the migration has completed, or `False` otherwise.
- `ImportTime` – Timestamp.

  The time that the migration was started.
- `ImportedBy` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the person who initiated the migration.

Operations

- ImportCatalogToGlue action (Python: import_catalog_to_glue)
- GetCatalogImportStatus action (Python: get_catalog_import_status)

ImportCatalogToGlue action (Python: import_catalog_to_glue)

Imports an existing Amazon Athena Data Catalog to AWS Glue.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the catalog to import. Currently, this should be the AWS account ID.

Response

- *No Response parameters.*

Errors

- `InternalServiceException`
- `OperationTimeoutException`

GetCatalogImportStatus action (Python: get_catalog_import_status)

Retrieves the status of a migration operation.

Request

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the catalog to migrate. Currently, this should be the AWS account ID.

Response

- `ImportStatus` – A CatalogImportStatus object.

  The status of the specified catalog migration.

Errors

- `InternalServiceException`
- `OperationTimeoutException`

Table optimizer API

The table optimizer API describes the AWS Glue API for enabling compaction to improve read performance.

Data types

- TableOptimizer structure
- TableOptimizerConfiguration structure
- TableOptimizerVpcConfiguration structure
- CompactionConfiguration structure
- IcebergCompactionConfiguration structure
- TableOptimizerRun structure
- BatchGetTableOptimizerEntry structure
- BatchTableOptimizer structure
- BatchGetTableOptimizerError structure
- RetentionConfiguration structure
- IcebergRetentionConfiguration structure
- OrphanFileDeletionConfiguration structure
- IcebergOrphanFileDeletionConfiguration structure
- CompactionMetrics structure
- RetentionMetrics structure
- OrphanFileDeletionMetrics structure
- IcebergCompactionMetrics structure
- IcebergRetentionMetrics structure
- IcebergOrphanFileDeletionMetrics structure
- RunMetrics structure

TableOptimizer structure

Contains details about an optimizer associated with a table.

Fields

- `type` – UTF-8 string (valid values: `compaction="COMPACTION"` | `retention="RETENTION"` | `orphan_file_deletion="ORPHAN_FILE_DELETION"`).

The type of table optimizer. The valid values are:

- `compaction`: for managing compaction with a table optimizer.
- `retention`: for managing the retention of snapshot with a table optimizer.
- `orphan_file_deletion`: for managing the deletion of orphan files with a table optimizer.

- `configuration` – A TableOptimizerConfiguration object.

  A `TableOptimizerConfiguration` object that was specified when creating or updating a table optimizer.
- `lastRun` – A TableOptimizerRun object.

  A `TableOptimizerRun` object representing the last run of the table optimizer.
- `configurationSource` – UTF-8 string (valid values: `catalog="CATALOG"` | `table="TABLE"`).

  Specifies the source of the optimizer configuration. This indicates how the table optimizer was configured and which entity or service initiated the configuration.

TableOptimizerConfiguration structure

Contains details on the configuration of a table optimizer. You pass this configuration when creating or updating a table optimizer.

Fields

- `roleArn` – UTF-8 string, not less than 20 or more than 2048 bytes long, matching the Single-line string pattern.

  A role passed by the caller which gives the service permission to update the resources associated with the optimizer on the caller's behalf.
- `enabled` – Boolean.

  Whether table optimization is enabled.
- `vpcConfiguration` – A TableOptimizerVpcConfiguration object.

  A `TableOptimizerVpcConfiguration` object representing the VPC configuration for a table optimizer.

  This configuration is necessary to perform optimization on tables that are in a customer VPC.

- `compactionConfiguration` – A CompactionConfiguration object.

  The configuration for a compaction optimizer. This configuration defines how data files in your table will be compacted to improve query performance and reduce storage costs.
- `retentionConfiguration` – A RetentionConfiguration object.

  The configuration for a snapshot retention optimizer.
- `orphanFileDeletionConfiguration` – An OrphanFileDeletionConfiguration object.

  The configuration for an orphan file deletion optimizer.

TableOptimizerVpcConfiguration structure

An object that describes the VPC configuration for a table optimizer.

This configuration is necessary to perform optimization on tables that are in a customer VPC.

Fields

- `glueConnectionName` – UTF-8 string, at least 1 byte long.

  The name of the AWS Glue connection used for the VPC for the table optimizer.

CompactionConfiguration structure

The configuration for a compaction optimizer. This configuration defines how data files in your table will be compacted to improve query performance and reduce storage costs.

Fields

- `icebergConfiguration` – An IcebergCompactionConfiguration object.

  The configuration for an Iceberg compaction optimizer.

IcebergCompactionConfiguration structure

The configuration for an Iceberg compaction optimizer. This configuration defines parameters for optimizing the layout of data files in Iceberg tables.

Fields

- `strategy` – UTF-8 string (valid values: `binpack="BINPACK"` | `sort="SORT"` | `z-order="ZORDER"`).

  The strategy to use for compaction. Valid values are:

  - `binpack`: Combines small files into larger files, typically targeting sizes over 100MB, while applying any pending deletes. This is the recommended compaction strategy for most use cases.
  - `sort`: Organizes data based on specified columns which are sorted hierarchically during compaction, improving query performance for filtered operations. This strategy is recommended when your queries frequently filter on specific columns. To use this strategy, you must first define a sort order in your Iceberg table properties using the `sort_order` table property.
  - `z-order`: Optimizes data organization by blending multiple attributes into a single scalar value that can be used for sorting, allowing efficient querying across multiple dimensions. This strategy is recommended when you need to query data across multiple dimensions simultaneously. To use this strategy, you must first define a sort order in your Iceberg table properties using the `sort_order` table property.

  If an input is not provided, the default value 'binpack' will be used.

- `minInputFiles` – Number (integer).

  The minimum number of data files that must be present in a partition before compaction will actually compact files. This parameter helps control when compaction is triggered, preventing unnecessary compaction operations on partitions with few files. If an input is not provided, the default value 100 will be used.

- `deleteFileThreshold` – Number (integer).

  The minimum number of deletes that must be present in a data file to make it eligible for compaction. This parameter helps optimize compaction by focusing on files that contain a significant number of delete operations, which can improve query performance by removing deleted records. If an input is not provided, the default value 1 will be used.

TableOptimizerRun structure

Contains details for a table optimizer run.

Fields

- `eventType` – UTF-8 string (valid values: `starting="STARTING"` | `completed="COMPLETED"` | `failed="FAILED"` | `in_progress="IN_PROGRESS"`).

  An event type representing the status of the table optimizer run.
- `startTimestamp` – Timestamp.

  Represents the epoch timestamp at which the compaction job was started within Lake Formation.
- `endTimestamp` – Timestamp.

  Represents the epoch timestamp at which the compaction job ended.
- `metrics` – A RunMetrics object.

  A `RunMetrics` object containing metrics for the optimizer run.

  This member is deprecated. See the individual metric members for compaction, retention, and orphan file deletion.
- `error` – UTF-8 string.

  An error that occured during the optimizer run.
- `compactionMetrics` – A CompactionMetrics object.

  A `CompactionMetrics` object containing metrics for the optimizer run.
- `compactionStrategy` – UTF-8 string (valid values: `binpack="BINPACK"` | `sort="SORT"` | `z-order="ZORDER"`).

  The strategy used for the compaction run. Indicates which algorithm was applied to determine how files were selected and combined during the compaction process. Valid values are:
  - `binpack`: Combines small files into larger files, typically targeting sizes over 100MB, while applying any pending deletes. This is the recommended compaction strategy for most use cases.
  - `sort`: Organizes data based on specified columns which are sorted hierarchically during compaction, improving query performance for filtered operations. This strategy is recommended when your queries frequently filter on specific columns. To use this strategy, you must first define a sort order in your Iceberg table properties using the `sort_order` table property.

  - `z-order`: Optimizes data organization by blending multiple attributes into a single scalar value that can be used for sorting, allowing efficient querying across multiple dimensions. This strategy is recommended when you need to query data across multiple dimensions simultaneously. To use this strategy, you must first define a sort order in your Iceberg table properties using the `sort_order` table property.
- `retentionMetrics` – A RetentionMetrics object.

  A `RetentionMetrics` object containing metrics for the optimizer run.
- `orphanFileDeletionMetrics` – An OrphanFileDeletionMetrics object.

  An `OrphanFileDeletionMetrics` object containing metrics for the optimizer run.

BatchGetTableOptimizerEntry structure

Represents a table optimizer to retrieve in the `BatchGetTableOptimizer` operation.

Fields

- `catalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Catalog ID of the table.
- `databaseName` – UTF-8 string, at least 1 byte long.

  The name of the database in the catalog in which the table resides.
- `tableName` – UTF-8 string, at least 1 byte long.

  The name of the table.
- `type` – UTF-8 string (valid values: `compaction="COMPACTION"` | `retention="RETENTION"` | `orphan_file_deletion="ORPHAN_FILE_DELETION"`).

  The type of table optimizer.

BatchTableOptimizer structure

Contains details for one of the table optimizers returned by the `BatchGetTableOptimizer` operation.

Fields

- `catalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Catalog ID of the table.
- `databaseName` – UTF-8 string, at least 1 byte long.

  The name of the database in the catalog in which the table resides.
- `tableName` – UTF-8 string, at least 1 byte long.

  The name of the table.
- `tableOptimizer` – A TableOptimizer object.

  A `TableOptimizer` object that contains details on the configuration and last run of a table optimizer.

BatchGetTableOptimizerError structure

Contains details on one of the errors in the error list returned by the `BatchGetTableOptimizer` operation.

Fields

- `error` – An ErrorDetail object.

  An `ErrorDetail` object containing code and message details about the error.
- `catalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Catalog ID of the table.
- `databaseName` – UTF-8 string, at least 1 byte long.

  The name of the database in the catalog in which the table resides.
- `tableName` – UTF-8 string, at least 1 byte long.

  The name of the table.
- `type` – UTF-8 string (valid values: `compaction="COMPACTION"` | `retention="RETENTION"` | `orphan_file_deletion="ORPHAN_FILE_DELETION"`).

The type of table optimizer.

RetentionConfiguration structure

The configuration for a snapshot retention optimizer.

Fields

- `icebergConfiguration` – An IcebergRetentionConfiguration object.

  The configuration for an Iceberg snapshot retention optimizer.

IcebergRetentionConfiguration structure

The configuration for an Iceberg snapshot retention optimizer.

Fields

- `snapshotRetentionPeriodInDays` – Number (integer).

  The number of days to retain the Iceberg snapshots. If an input is not provided, the corresponding Iceberg table configuration field will be used or if not present, the default value 5 will be used.
- `numberOfSnapshotsToRetain` – Number (integer).

  The number of Iceberg snapshots to retain within the retention period. If an input is not provided, the corresponding Iceberg table configuration field will be used or if not present, the default value 1 will be used.
- `cleanExpiredFiles` – Boolean.

  If set to false, snapshots are only deleted from table metadata, and the underlying data and metadata files are not deleted.
- `runRateInHours` – Number (integer).

  The interval in hours between retention job runs. This parameter controls how frequently the retention optimizer will run to clean up expired snapshots. The value must be between 3 and 168 hours (7 days). If an input is not provided, the default value 24 will be used.

OrphanFileDeletionConfiguration structure

The configuration for an orphan file deletion optimizer.

Fields

- `icebergConfiguration` – An IcebergOrphanFileDeletionConfiguration object.

  The configuration for an Iceberg orphan file deletion optimizer.

IcebergOrphanFileDeletionConfiguration structure

The configuration for an Iceberg orphan file deletion optimizer.

Fields

- `orphanFileRetentionPeriodInDays` – Number (integer).

  The number of days that orphan files should be retained before file deletion. If an input is not provided, the default value 3 will be used.
- `location` – UTF-8 string.

  Specifies a directory in which to look for files (defaults to the table's location). You may choose a sub-directory rather than the top-level table location.
- `runRateInHours` – Number (integer).

  The interval in hours between orphan file deletion job runs. This parameter controls how frequently the orphan file deletion optimizer will run to clean up orphan files. The value must be between 3 and 168 hours (7 days). If an input is not provided, the default value 24 will be used.

CompactionMetrics structure

A structure that contains compaction metrics for the optimizer run.

Fields

- `IcebergMetrics` – An IcebergCompactionMetrics object.

  A structure containing the Iceberg compaction metrics for the optimizer run.

RetentionMetrics structure

A structure that contains retention metrics for the optimizer run.

Fields

- `IcebergMetrics` – An IcebergRetentionMetrics object.

  A structure containing the Iceberg retention metrics for the optimizer run.

OrphanFileDeletionMetrics structure

A structure that contains orphan file deletion metrics for the optimizer run.

Fields

- `IcebergMetrics` – An IcebergOrphanFileDeletionMetrics object.

  A structure containing the Iceberg orphan file deletion metrics for the optimizer run.

IcebergCompactionMetrics structure

Compaction metrics for Iceberg for the optimizer run.

Fields

- `DpuHours` – Number (double).

  The number of DPU hours consumed by the job.
- `NumberOfDpus` – Number (Integer).

  The number of DPUs consumed by the job, rounded up to the nearest whole number.
- `JobDurationInHour` – Number (double).

  The duration of the job in hours.

IcebergRetentionMetrics structure

Snapshot retention metrics for Iceberg for the optimizer run.

Fields

- `DpuHours` – Number (double).

  The number of DPU hours consumed by the job.
- `NumberOfDpus` – Number (Integer).

  The number of DPUs consumed by the job, rounded up to the nearest whole number.
- `JobDurationInHour` – Number (double).

  The duration of the job in hours.

IcebergOrphanFileDeletionMetrics structure

Orphan file deletion metrics for Iceberg for the optimizer run.

Fields

- `DpuHours` – Number (double).

  The number of DPU hours consumed by the job.
- `NumberOfDpus` – Number (Integer).

  The number of DPUs consumed by the job, rounded up to the nearest whole number.
- `JobDurationInHour` – Number (double).

  The duration of the job in hours.

RunMetrics structure

Metrics for the optimizer run.

This structure is deprecated. See the individual metric members for compaction, retention, and orphan file deletion.

Fields

- `NumberOfBytesCompacted` – UTF-8 string.

  The number of bytes removed by the compaction job run.

- `NumberOfFilesCompacted` – UTF-8 string.

  The number of files removed by the compaction job run.
- `NumberOfDpus` – UTF-8 string.

  The number of DPUs consumed by the job, rounded up to the nearest whole number.
- `JobDurationInHour` – UTF-8 string.

  The duration of the job in hours.

Operations

- GetTableOptimizer action (Python: get_table_optimizer)
- BatchGetTableOptimizer action (Python: batch_get_table_optimizer)
- ListTableOptimizerRuns action (Python: list_table_optimizer_runs)
- CreateTableOptimizer action (Python: create_table_optimizer)
- DeleteTableOptimizer action (Python: delete_table_optimizer)
- UpdateTableOptimizer action (Python: update_table_optimizer)

GetTableOptimizer action (Python: get_table_optimizer)

Returns the configuration of all optimizers associated with a specified table.

Request

- `CatalogId` – *Required:* Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Catalog ID of the table.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the database in the catalog in which the table resides.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table.

- `Type` – *Required:* UTF-8 string (valid values: `compaction="COMPACTION"` | `retention="RETENTION"` | `orphan_file_deletion="ORPHAN_FILE_DELETION"`).

  The type of table optimizer.

Response

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Catalog ID of the table.
- `DatabaseName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the database in the catalog in which the table resides.
- `TableName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table.
- `TableOptimizer` – A TableOptimizer object.

  The optimizer associated with the specified table.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `AccessDeniedException`
- `InternalServiceException`
- `ThrottlingException`

BatchGetTableOptimizer action (Python: batch_get_table_optimizer)

Returns the configuration for the specified table optimizers.

Request

- `Entries` – *Required:* An array of BatchGetTableOptimizerEntry objects.

  A list of `BatchGetTableOptimizerEntry` objects specifying the table optimizers to retrieve.

Response

- `TableOptimizers` – An array of BatchTableOptimizer objects.

  A list of `BatchTableOptimizer` objects.
- `Failures` – An array of BatchGetTableOptimizerError objects.

  A list of errors from the operation.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `AccessDeniedException`
- `InternalServiceException`
- `ThrottlingException`

ListTableOptimizerRuns action (Python: list_table_optimizer_runs)

Lists the history of previous optimizer runs for a specific table.

Request

- `CatalogId` – *Required:* Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Catalog ID of the table.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the database in the catalog in which the table resides.

- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table.
- `Type` – *Required:* UTF-8 string (valid values: `compaction="COMPACTION"` | `retention="RETENTION"` | `orphan_file_deletion="ORPHAN_FILE_DELETION"`).

  The type of table optimizer.
- `MaxResults` – Number (integer).

  The maximum number of optimizer runs to return on each call.
- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation call.

Response

- `CatalogId` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Catalog ID of the table.
- `DatabaseName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the database in the catalog in which the table resides.
- `TableName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table.
- `NextToken` – UTF-8 string.

  A continuation token for paginating the returned list of optimizer runs, returned if the current segment of the list is not the last.
- `TableOptimizerRuns` – An array of TableOptimizerRun objects.

  A list of the optimizer runs associated with a table.

Errors

- `EntityNotFoundException`
- `AccessDeniedException`
- `InvalidInputException`
- `ValidationException`
- `InternalServiceException`
- `ThrottlingException`

CreateTableOptimizer action (Python: create_table_optimizer)

Creates a new table optimizer for a specific function.

Request

- `CatalogId` – *Required:* Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Catalog ID of the table.

- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the database in the catalog in which the table resides.

- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table.

- `Type` – *Required:* UTF-8 string (valid values: `compaction="COMPACTION"` | `retention="RETENTION"` | `orphan_file_deletion="ORPHAN_FILE_DELETION"`).

  The type of table optimizer.

- `TableOptimizerConfiguration` – *Required:* A TableOptimizerConfiguration object.

  A `TableOptimizerConfiguration` object representing the configuration of a table optimizer.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `ValidationException`
- `InvalidInputException`
- `AccessDeniedException`
- `AlreadyExistsException`
- `InternalServiceException`
- `ThrottlingException`

DeleteTableOptimizer action (Python: delete_table_optimizer)

Deletes an optimizer and all associated metadata for a table. The optimization will no longer be performed on the table.

Request

- `CatalogId` – *Required:* Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Catalog ID of the table.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the database in the catalog in which the table resides.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table.
- `Type` – *Required:* UTF-8 string (valid values: `compaction="COMPACTION"` | `retention="RETENTION"` | `orphan_file_deletion="ORPHAN_FILE_DELETION"`).

  The type of table optimizer.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `AccessDeniedException`
- `InternalServiceException`
- `ThrottlingException`

UpdateTableOptimizer action (Python: update_table_optimizer)

Updates the configuration for an existing table optimizer.

Request

- `CatalogId` – *Required:* Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Catalog ID of the table.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the database in the catalog in which the table resides.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table.
- `Type` – *Required:* UTF-8 string (valid values: `compaction="COMPACTION"` | `retention="RETENTION"` | `orphan_file_deletion="ORPHAN_FILE_DELETION"`).

  The type of table optimizer.
- `TableOptimizerConfiguration` – *Required:* A TableOptimizerConfiguration object.

  A `TableOptimizerConfiguration` object representing the configuration of a table optimizer.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `AccessDeniedException`
- `ValidationException`
- `InternalServiceException`
- `ThrottlingException`
- `ConcurrentModificationException`

Crawlers and classifiers API

The Crawler and classifiers API describes the AWS Glue crawler and classifier data types, and includes the API for creating, deleting, updating, and listing crawlers or classifiers.

Topics

- Classifier API
- Crawler API
- Column statistics API
- Crawler scheduler API

Classifier API

The Classifier API describes AWS Glue classifier data types, and includes the API for creating, deleting, updating, and listing classifiers.

Data types

- Classifier structure
- GrokClassifier structure
- XMLClassifier structure

- JsonClassifier structure
- CsvClassifier structure
- CreateGrokClassifierRequest structure
- UpdateGrokClassifierRequest structure
- CreateXMLClassifierRequest structure
- UpdateXMLClassifierRequest structure
- CreateJsonClassifierRequest structure
- UpdateJsonClassifierRequest structure
- CreateCsvClassifierRequest structure
- UpdateCsvClassifierRequest structure

Classifier structure

Classifiers are triggered during a crawl task. A classifier checks whether a given file is in a format it can handle. If it is, the classifier creates a schema in the form of a `StructType` object that matches that data format.

You can use the standard classifiers that AWS Glue provides, or you can write your own classifiers to best categorize your data sources and specify the appropriate schemas to use for them. A classifier can be a `grok` classifier, an `XML` classifier, a `JSON` classifier, or a custom `CSV` classifier, as specified in one of the fields in the `Classifier` object.

Fields

- `GrokClassifier` – A GrokClassifier object.

  A classifier that uses `grok`.
- `XMLClassifier` – A XMLClassifier object.

  A classifier for XML content.
- `JsonClassifier` – A JsonClassifier object.

  A classifier for JSON content.
- `CsvClassifier` – A CsvClassifier object.

  A classifier for comma-separated values (CSV).

GrokClassifier structure

A classifier that uses `grok` patterns.

Fields

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the classifier.
- `Classification` – *Required:* UTF-8 string.

  An identifier of the data format that the classifier matches, such as Twitter, JSON, Omniture logs, and so on.
- `CreationTime` – Timestamp.

  The time that this classifier was registered.
- `LastUpdated` – Timestamp.

  The time that this classifier was last updated.
- `Version` – Number (long).

  The version of this classifier.
- `GrokPattern` – *Required:* UTF-8 string, not less than 1 or more than 2048 bytes long, matching the A Logstash Grok string pattern.

  The grok pattern applied to a data store by this classifier. For more information, see built-in patterns in Writing Custom Classifiers.
- `CustomPatterns` – UTF-8 string, not more than 16000 bytes long, matching the URI address multi-line string pattern.

  Optional custom grok patterns defined by this classifier. For more information, see custom patterns in Writing Custom Classifiers.

XMLClassifier structure

A classifier for `XML` content.

Fields

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the classifier.
- `Classification` – *Required:* UTF-8 string.

  An identifier of the data format that the classifier matches.
- `CreationTime` – Timestamp.

  The time that this classifier was registered.
- `LastUpdated` – Timestamp.

  The time that this classifier was last updated.
- `Version` – Number (long).

  The version of this classifier.
- `RowTag` – UTF-8 string.

  The XML tag designating the element that contains each record in an XML document being parsed. This can't identify a self-closing element (closed by `/>`). An empty row element that contains only attributes can be parsed as long as it ends with a closing tag (for example, `<row item_a="A" item_b="B"></row>` is okay, but `<row item_a="A" item_b="B" />` is not).

JsonClassifier structure

A classifier for `JSON` content.

Fields

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the classifier.
- `CreationTime` – Timestamp.

  The time that this classifier was registered.
- `LastUpdated` – Timestamp.

The time that this classifier was last updated.

- `Version` – Number (long).

  The version of this classifier.
- `JsonPath` – *Required:* UTF-8 string.

  A `JsonPath` string defining the JSON data for the classifier to classify. AWS Glue supports a subset of JsonPath, as described in Writing JsonPath Custom Classifiers.

CsvClassifier structure

A classifier for custom `CSV` content.

Fields

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the classifier.
- `CreationTime` – Timestamp.

  The time that this classifier was registered.
- `LastUpdated` – Timestamp.

  The time that this classifier was last updated.
- `Version` – Number (long).

  The version of this classifier.
- `Delimiter` – UTF-8 string, not less than 1 or more than 1 bytes long, matching the Custom string pattern #26.

  A custom symbol to denote what separates each column entry in the row.
- `QuoteSymbol` – UTF-8 string, not less than 1 or more than 1 bytes long, matching the Custom string pattern #26.

  A custom symbol to denote what combines content into a single column value. It must be different from the column delimiter.
- `ContainsHeader` – UTF-8 string (valid values: `UNKNOWN` | `PRESENT` | `ABSENT`).

Indicates whether the CSV file contains a header.

- `Header` – An array of UTF-8 strings.

  A list of strings representing column names.
- `DisableValueTrimming` – Boolean.

  Specifies not to trim values before identifying the type of column values. The default value is `true`.
- `AllowSingleColumn` – Boolean.

  Enables the processing of files that contain only one column.
- `CustomDatatypeConfigured` – Boolean.

  Enables the custom datatype to be configured.
- `CustomDatatypes` – An array of UTF-8 strings.

  A list of custom datatypes including "BINARY", "BOOLEAN", "DATE", "DECIMAL", "DOUBLE", "FLOAT", "INT", "LONG", "SHORT", "STRING", "TIMESTAMP".
- `Serde` – UTF-8 string (valid values: `OpenCSVSerDe` | `LazySimpleSerDe` | `None`).

  Sets the SerDe for processing CSV in the classifier, which will be applied in the Data Catalog. Valid values are `OpenCSVSerDe`, `LazySimpleSerDe`, and `None`. You can specify the `None` value when you want the crawler to do the detection.

CreateGrokClassifierRequest structure

Specifies a `grok` classifier for `CreateClassifier` to create.

Fields

- `Classification` – *Required:* UTF-8 string.

  An identifier of the data format that the classifier matches, such as Twitter, JSON, Omniture logs, Amazon CloudWatch Logs, and so on.
- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the [Single-line string pattern](#).

  The name of the new classifier.

- `GrokPattern` – *Required:* UTF-8 string, not less than 1 or more than 2048 bytes long, matching the A Logstash Grok string pattern.

  The grok pattern used by this classifier.
- `CustomPatterns` – UTF-8 string, not more than 16000 bytes long, matching the URI address multi-line string pattern.

  Optional custom grok patterns used by this classifier.

UpdateGrokClassifierRequest structure

Specifies a grok classifier to update when passed to `UpdateClassifier`.

Fields

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the `GrokClassifier`.
- `Classification` – UTF-8 string.

  An identifier of the data format that the classifier matches, such as Twitter, JSON, Omniture logs, Amazon CloudWatch Logs, and so on.
- `GrokPattern` – UTF-8 string, not less than 1 or more than 2048 bytes long, matching the A Logstash Grok string pattern.

  The grok pattern used by this classifier.
- `CustomPatterns` – UTF-8 string, not more than 16000 bytes long, matching the URI address multi-line string pattern.

  Optional custom grok patterns used by this classifier.

CreateXMLClassifierRequest structure

Specifies an XML classifier for `CreateClassifier` to create.

Fields

- `Classification` – *Required:* UTF-8 string.

An identifier of the data format that the classifier matches.

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the classifier.

- `RowTag` – UTF-8 string.

  The XML tag designating the element that contains each record in an XML document being parsed. This can't identify a self-closing element (closed by `/>`). An empty row element that contains only attributes can be parsed as long as it ends with a closing tag (for example, `<row item_a="A" item_b="B"></row>` is okay, but `<row item_a="A" item_b="B" />` is not).

UpdateXMLClassifierRequest structure

Specifies an XML classifier to be updated.

Fields

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the classifier.

- `Classification` – UTF-8 string.

  An identifier of the data format that the classifier matches.

- `RowTag` – UTF-8 string.

  The XML tag designating the element that contains each record in an XML document being parsed. This cannot identify a self-closing element (closed by `/>`). An empty row element that contains only attributes can be parsed as long as it ends with a closing tag (for example, `<row item_a="A" item_b="B"></row>` is okay, but `<row item_a="A" item_b="B" />` is not).

CreateJsonClassifierRequest structure

Specifies a JSON classifier for `CreateClassifier` to create.

Fields

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the classifier.
- `JsonPath` – *Required:* UTF-8 string.

  A `JsonPath` string defining the JSON data for the classifier to classify. AWS Glue supports a subset of JsonPath, as described in Writing JsonPath Custom Classifiers.

UpdateJsonClassifierRequest structure

Specifies a JSON classifier to be updated.

Fields

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the classifier.
- `JsonPath` – UTF-8 string.

  A `JsonPath` string defining the JSON data for the classifier to classify. AWS Glue supports a subset of JsonPath, as described in Writing JsonPath Custom Classifiers.

CreateCsvClassifierRequest structure

Specifies a custom CSV classifier for `CreateClassifier` to create.

Fields

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the classifier.
- `Delimiter` – UTF-8 string, not less than 1 or more than 1 bytes long, matching the Custom string pattern #26.

  A custom symbol to denote what separates each column entry in the row.

- `QuoteSymbol` – UTF-8 string, not less than 1 or more than 1 bytes long, matching the Custom string pattern #26.

  A custom symbol to denote what combines content into a single column value. Must be different from the column delimiter.
- `ContainsHeader` – UTF-8 string (valid values: `UNKNOWN` | `PRESENT` | `ABSENT`).

  Indicates whether the CSV file contains a header.
- `Header` – An array of UTF-8 strings.

  A list of strings representing column names.
- `DisableValueTrimming` – Boolean.

  Specifies not to trim values before identifying the type of column values. The default value is true.
- `AllowSingleColumn` – Boolean.

  Enables the processing of files that contain only one column.
- `CustomDatatypeConfigured` – Boolean.

  Enables the configuration of custom datatypes.
- `CustomDatatypes` – An array of UTF-8 strings.

  Creates a list of supported custom datatypes.
- `Serde` – UTF-8 string (valid values: `OpenCSVSerDe` | `LazySimpleSerDe` | `None`).

  Sets the SerDe for processing CSV in the classifier, which will be applied in the Data Catalog. Valid values are `OpenCSVSerDe`, `LazySimpleSerDe`, and `None`. You can specify the `None` value when you want the crawler to do the detection.

UpdateCsvClassifierRequest structure

Specifies a custom CSV classifier to be updated.

Fields

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The name of the classifier.

- `Delimiter` – UTF-8 string, not less than 1 or more than 1 bytes long, matching the Custom string pattern #26.

  A custom symbol to denote what separates each column entry in the row.
- `QuoteSymbol` – UTF-8 string, not less than 1 or more than 1 bytes long, matching the Custom string pattern #26.

  A custom symbol to denote what combines content into a single column value. It must be different from the column delimiter.
- `ContainsHeader` – UTF-8 string (valid values: `UNKNOWN` | `PRESENT` | `ABSENT`).

  Indicates whether the CSV file contains a header.
- `Header` – An array of UTF-8 strings.

  A list of strings representing column names.
- `DisableValueTrimming` – Boolean.

  Specifies not to trim values before identifying the type of column values. The default value is true.
- `AllowSingleColumn` – Boolean.

  Enables the processing of files that contain only one column.
- `CustomDatatypeConfigured` – Boolean.

  Specifies the configuration of custom datatypes.
- `CustomDatatypes` – An array of UTF-8 strings.

  Specifies a list of supported custom datatypes.
- `Serde` – UTF-8 string (valid values: `OpenCSVSerDe` | `LazySimpleSerDe` | `None`).

  Sets the SerDe for processing CSV in the classifier, which will be applied in the Data Catalog. Valid values are `OpenCSVSerDe`, `LazySimpleSerDe`, and `None`. You can specify the `None` value when you want the crawler to do the detection.

Operations

- CreateClassifier action (Python: create_classifier)
- DeleteClassifier action (Python: delete_classifier)
- GetClassifier action (Python: get_classifier)
- GetClassifiers action (Python: get_classifiers)
- UpdateClassifier action (Python: update_classifier)

CreateClassifier action (Python: create_classifier)

Creates a classifier in the user's account. This can be a `GrokClassifier`, an `XMLClassifier`, a `JsonClassifier`, or a `CsvClassifier`, depending on which field of the request is present.

Request

- `GrokClassifier` – A CreateGrokClassifierRequest object.

  A `GrokClassifier` object specifying the classifier to create.
- `XMLClassifier` – A CreateXMLClassifierRequest object.

  An `XMLClassifier` object specifying the classifier to create.
- `JsonClassifier` – A CreateJsonClassifierRequest object.

  A `JsonClassifier` object specifying the classifier to create.
- `CsvClassifier` – A CreateCsvClassifierRequest object.

  A `CsvClassifier` object specifying the classifier to create.

Response

- *No Response parameters.*

Errors

- `AlreadyExistsException`
- `InvalidInputException`
- `OperationTimeoutException`

DeleteClassifier action (Python: delete_classifier)

Removes a classifier from the Data Catalog.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Name of the classifier to remove.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `OperationTimeoutException`

GetClassifier action (Python: get_classifier)

Retrieve a classifier by name.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Name of the classifier to retrieve.

Response

- `Classifier` – A Classifier object.

  The requested classifier.

Errors

- `EntityNotFoundException`
- `OperationTimeoutException`

GetClassifiers action (Python: get_classifiers)

Lists all classifier objects in the Data Catalog.

Request

- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The size of the list to return (optional).
- `NextToken` – UTF-8 string.

  An optional continuation token.

Response

- `Classifiers` – An array of Classifier objects.

  The requested list of classifier objects.
- `NextToken` – UTF-8 string.

  A continuation token.

Errors

- `OperationTimeoutException`

UpdateClassifier action (Python: update_classifier)

Modifies an existing classifier (a `GrokClassifier`, an `XMLClassifier`, a `JsonClassifier`, or a `CsvClassifier`, depending on which field is present).

Request

- `GrokClassifier` – An UpdateGrokClassifierRequest object.

A `GrokClassifier` object with updated fields.

- `XMLClassifier` – An UpdateXMLClassifierRequest object.

  An `XMLClassifier` object with updated fields.
- `JsonClassifier` – An UpdateJsonClassifierRequest object.

  A `JsonClassifier` object with updated fields.
- `CsvClassifier` – An UpdateCsvClassifierRequest object.

  A `CsvClassifier` object with updated fields.

Response

- *No Response parameters.*

Errors

- `InvalidInputException`
- `VersionMismatchException`
- `EntityNotFoundException`
- `OperationTimeoutException`

Crawler API

The Crawler API describes AWS Glue crawler data types, along with the API for creating, deleting, updating, and listing crawlers.

Data types

- Crawler structure
- Schedule structure
- CrawlerTargets structure
- S3Target structure
- S3DeltaCatalogTarget structure
- S3DeltaDirectTarget structure

- JdbcTarget structure
- MongoDBTarget structure
- DynamoDBTarget structure
- DeltaTarget structure
- IcebergTarget structure
- HudiTarget structure
- CatalogTarget structure
- CrawlerMetrics structure
- CrawlerHistory structure
- CrawlsFilter structure
- SchemaChangePolicy structure
- LastCrawlInfo structure
- RecrawlPolicy structure
- LineageConfiguration structure
- LakeFormationConfiguration structure

Crawler structure

Specifies a crawler program that examines a data source and uses classifiers to try to determine its schema. If successful, the crawler records metadata concerning the data source in the AWS Glue Data Catalog.

Fields

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the crawler.
- `Role` – UTF-8 string.

  The Amazon Resource Name (ARN) of an IAM role that's used to access customer resources, such as Amazon Simple Storage Service (Amazon S3) data.
- `Targets` – A CrawlerTargets object.

A collection of targets to crawl.

- `DatabaseName` – UTF-8 string.

  The name of the database in which the crawler's output is stored.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the crawler.
- `Classifiers` – An array of UTF-8 strings.

  A list of UTF-8 strings that specify the custom classifiers that are associated with the crawler.
- `RecrawlPolicy` – A RecrawlPolicy object.

  A policy that specifies whether to crawl the entire dataset again, or to crawl only folders that were added since the last crawler run.
- `SchemaChangePolicy` – A SchemaChangePolicy object.

  The policy that specifies update and delete behaviors for the crawler.
- `LineageConfiguration` – A LineageConfiguration object.

  A configuration that specifies whether data lineage is enabled for the crawler.
- `State` – UTF-8 string (valid values: `READY` | `RUNNING` | `STOPPING`).

  Indicates whether the crawler is running, or whether a run is pending.
- `TablePrefix` – UTF-8 string, not more than 128 bytes long.

  The prefix added to the names of tables that are created.
- `Schedule` – A Schedule object.

  For scheduled crawlers, the schedule when the crawler runs.
- `CrawlElapsedTime` – Number (long).

  If the crawler is running, contains the total time elapsed since the last crawl began.
- `CreationTime` – Timestamp.

  The time that the crawler was created.
- `LastUpdated` – Timestamp.

The time that the crawler was last updated.

- `LastCrawl` – A LastCrawlInfo object.

  The status of the last crawl, and potentially error information if an error occurred.
- `Version` – Number (long).

  The version of the crawler.
- `Configuration` – UTF-8 string.

  Crawler configuration information. This versioned JSON string allows users to specify aspects of a crawler's behavior. For more information, see Setting crawler configuration options.
- `CrawlerSecurityConfiguration` – UTF-8 string, not more than 128 bytes long.

  The name of the `SecurityConfiguration` structure to be used by this crawler.
- `LakeFormationConfiguration` – A LakeFormationConfiguration object.

  Specifies whether the crawler should use AWS Lake Formation credentials for the crawler instead of the IAM role credentials.

Schedule structure

A scheduling object using a `cron` statement to schedule an event.

Fields

- `ScheduleExpression` – UTF-8 string.

  A `cron` expression used to specify the schedule (see Time-Based Schedules for Jobs and Crawlers. For example, to run something every day at 12:15 UTC, you would specify: `cron(15 12 * * ? *)`.
- `State` – UTF-8 string (valid values: `SCHEDULED` | `NOT_SCHEDULED` | `TRANSITIONING`).

  The state of the schedule.

CrawlerTargets structure

Specifies data stores to crawl.

Fields

- `S3Targets` – An array of [S3Target](#) objects.

  Specifies Amazon Simple Storage Service (Amazon S3) targets.
- `JdbcTargets` – An array of [JdbcTarget](#) objects.

  Specifies JDBC targets.
- `MongoDBTargets` – An array of [MongoDBTarget](#) objects.

  Specifies Amazon DocumentDB or MongoDB targets.
- `DynamoDBTargets` – An array of [DynamoDBTarget](#) objects.

  Specifies Amazon DynamoDB targets.
- `CatalogTargets` – An array of [CatalogTarget](#) objects.

  Specifies AWS Glue Data Catalog targets.
- `DeltaTargets` – An array of [DeltaTarget](#) objects.

  Specifies Delta data store targets.
- `IcebergTargets` – An array of [IcebergTarget](#) objects.

  Specifies Apache Iceberg data store targets.
- `HudiTargets` – An array of [HudiTarget](#) objects.

  Specifies Apache Hudi data store targets.

S3Target structure

Specifies a data store in Amazon Simple Storage Service (Amazon S3).

Fields

- `Path` – UTF-8 string.

  The path to the Amazon S3 target.
- `Exclusions` – An array of UTF-8 strings.

A list of glob patterns used to exclude from the crawl. For more information, see Catalog Tables with a Crawler.

- `ConnectionName` – UTF-8 string, not less than 1 or more than 2048 bytes long.

  The name of a connection which allows a job or crawler to access data in Amazon S3 within an Amazon Virtual Private Cloud environment (Amazon VPC).
- `SampleSize` – Number (integer).

  Sets the number of files in each leaf folder to be crawled when crawling sample files in a dataset. If not set, all the files are crawled. A valid value is an integer between 1 and 249.
- `EventQueueArn` – UTF-8 string.

  A valid Amazon SQS ARN. For example, `arn:aws:sqs:region:account:sqs`.
- `DlqEventQueueArn` – UTF-8 string.

  A valid Amazon dead-letter SQS ARN. For example, `arn:aws:sqs:region:account:deadLetterQueue`.

S3DeltaCatalogTarget structure

Specifies a target that writes to a Delta Lake data source in the AWS Glue Data Catalog.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data target.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The nodes that are inputs to the data target.
- `PartitionKeys` – An array of UTF-8 strings.

  Specifies native partitioning using a sequence of keys.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the table in the database to write to.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

The name of the database to write to.

- `AdditionalOptions` – A map array of key-value pairs.

  Each key is a UTF-8 string, matching the Custom string pattern #59.

  Each value is a UTF-8 string, matching the Custom string pattern #59.

  Specifies additional connection options for the connector.
- `SchemaChangePolicy` – A CatalogSchemaChangePolicy object.

  A policy that specifies update behavior for the crawler.
- `AutoDataQuality` – An AutoDataQuality object.

  Specifies whether to automatically enable data quality evaluation for the S3 Delta catalog target. When set to `true`, data quality checks are performed automatically during the write operation.
- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for the S3 Delta catalog target.

S3DeltaDirectTarget structure

Specifies a target that writes to a Delta Lake data source in Amazon S3.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data target.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The nodes that are inputs to the data target.
- `PartitionKeys` – An array of UTF-8 strings.

  Specifies native partitioning using a sequence of keys.
- `Path` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The Amazon S3 path of your Delta Lake data source to write to.

- `Compression` – *Required:* UTF-8 string (valid values: `uncompressed="UNCOMPRESSED"` | `snappy="SNAPPY"`).

  Specifies how the data is compressed. This is generally not necessary if the data has a standard file extension. Possible values are `"gzip"` and `"bzip"`).
- `NumberTargetPartitions` – UTF-8 string.

  Specifies the number of target partitions for distributing Delta Lake dataset files across Amazon S3.
- `Format` – *Required:* UTF-8 string (valid values: `json="JSON"` | `csv="CSV"` | `avro="AVRO"` | `orc="ORC"` | `parquet="PARQUET"` | `hudi="HUDI"` | `delta="DELTA"` | `iceberg="ICEBERG"` | `hyper="HYPER"` | `xml="XML"`).

  Specifies the data output format for the target.
- `AdditionalOptions` – A map array of key-value pairs.

  Each key is a UTF-8 string, matching the Custom string pattern #59.

  Each value is a UTF-8 string, matching the Custom string pattern #59.

  Specifies additional connection options for the connector.
- `SchemaChangePolicy` – A DirectSchemaChangePolicy object.

  A policy that specifies update behavior for the crawler.
- `AutoDataQuality` – An AutoDataQuality object.

  Specifies whether to automatically enable data quality evaluation for the S3 Delta direct target. When set to `true`, data quality checks are performed automatically during the write operation.

JdbcTarget structure

Specifies a JDBC data store to crawl.

Fields

- `ConnectionName` – UTF-8 string, not less than 1 or more than 2048 bytes long.

  The name of the connection to use to connect to the JDBC target.
- `Path` – UTF-8 string.

The path of the JDBC target.

- `Exclusions` – An array of UTF-8 strings.

  A list of glob patterns used to exclude from the crawl. For more information, see [Catalog Tables with a Crawler](#).
- `EnableAdditionalMetadata` – An array of UTF-8 strings.

  Specify a value of `RAWTYPES` or `COMMENTS` to enable additional metadata in table responses. `RAWTYPES` provides the native-level datatype. `COMMENTS` provides comments associated with a column or table in the database.

  If you do not need additional metadata, keep the field empty.

MongoDBTarget structure

Specifies an Amazon DocumentDB or MongoDB data store to crawl.

Fields

- `ConnectionName` – UTF-8 string, not less than 1 or more than 2048 bytes long.

  The name of the connection to use to connect to the Amazon DocumentDB or MongoDB target.
- `Path` – UTF-8 string.

  The path of the Amazon DocumentDB or MongoDB target (database/collection).
- `ScanAll` – Boolean.

  Indicates whether to scan all the records, or to sample rows from the table. Scanning all the records can take a long time when the table is not a high throughput table.

  A value of `true` means to scan all records, while a value of `false` means to sample the records. If no value is specified, the value defaults to `true`.

DynamoDBTarget structure

Specifies an Amazon DynamoDB table to crawl.

Fields

- `Path` – UTF-8 string.

  The name of the DynamoDB table to crawl.
- `scanAll` – Boolean.

  Indicates whether to scan all the records, or to sample rows from the table. Scanning all the records can take a long time when the table is not a high throughput table.

  A value of `true` means to scan all records, while a value of `false` means to sample the records. If no value is specified, the value defaults to `true`.
- `scanRate` – Number (double).

  The percentage of the configured read capacity units to use by the AWS Glue crawler. Read capacity units is a term defined by DynamoDB, and is a numeric value that acts as rate limiter for the number of reads that can be performed on that table per second.

  The valid values are null or a value between 0.1 to 1.5. A null value is used when user does not provide a value, and defaults to 0.5 of the configured Read Capacity Unit (for provisioned tables), or 0.25 of the max configured Read Capacity Unit (for tables using on-demand mode).

DeltaTarget structure

Specifies a Delta data store to crawl one or more Delta tables.

Fields

- `DeltaTables` – An array of UTF-8 strings.

  A list of the Amazon S3 paths to the Delta tables.
- `ConnectionName` – UTF-8 string, not less than 1 or more than 2048 bytes long.

  The name of the connection to use to connect to the Delta table target.
- `WriteManifest` – Boolean.

  Specifies whether to write the manifest files to the Delta table path.
- `CreateNativeDeltaTable` – Boolean.

Specifies whether the crawler will create native tables, to allow integration with query engines that support querying of the Delta transaction log directly.

IcebergTarget structure

Specifies an Apache Iceberg data source where Iceberg tables are stored in Amazon S3.

Fields

- `Paths` – An array of UTF-8 strings.

  One or more Amazon S3 paths that contains Iceberg metadata folders as `s3://bucket/prefix`.
- `ConnectionName` – UTF-8 string, not less than 1 or more than 2048 bytes long.

  The name of the connection to use to connect to the Iceberg target.
- `Exclusions` – An array of UTF-8 strings.

  A list of glob patterns used to exclude from the crawl. For more information, see Catalog Tables with a Crawler.
- `MaximumTraversalDepth` – Number (integer).

  The maximum depth of Amazon S3 paths that the crawler can traverse to discover the Iceberg metadata folder in your Amazon S3 path. Used to limit the crawler run time.

HudiTarget structure

Specifies an Apache Hudi data source.

Fields

- `Paths` – An array of UTF-8 strings.

  An array of Amazon S3 location strings for Hudi, each indicating the root folder with which the metadata files for a Hudi table resides. The Hudi folder may be located in a child folder of the root folder.

  The crawler will scan all folders underneath a path for a Hudi folder.

- `ConnectionName` – UTF-8 string, not less than 1 or more than 2048 bytes long.

  The name of the connection to use to connect to the Hudi target. If your Hudi files are stored in buckets that require VPC authorization, you can set their connection properties here.
- `Exclusions` – An array of UTF-8 strings.

  A list of glob patterns used to exclude from the crawl. For more information, see Catalog Tables with a Crawler.
- `MaximumTraversalDepth` – Number (integer).

  The maximum depth of Amazon S3 paths that the crawler can traverse to discover the Hudi metadata folder in your Amazon S3 path. Used to limit the crawler run time.

CatalogTarget structure

Specifies an AWS Glue Data Catalog target.

Fields

- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the database to be synchronized.
- `Tables` – *Required:* An array of UTF-8 strings, at least 1 string.

  A list of the tables to be synchronized.
- `ConnectionName` – UTF-8 string, not less than 1 or more than 2048 bytes long.

  The name of the connection for an Amazon S3-backed Data Catalog table to be a target of the crawl when using a `Catalog` connection type paired with a `NETWORK` Connection type.
- `EventQueueArn` – UTF-8 string.

  A valid Amazon SQS ARN. For example, `arn:aws:sqs:region:account:sqs`.
- `DlqEventQueueArn` – UTF-8 string.

  A valid Amazon dead-letter SQS ARN. For example, `arn:aws:sqs:region:account:deadLetterQueue`.

CrawlerMetrics structure

Metrics for a specified crawler.

Fields

- `CrawlerName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the crawler.
- `TimeLeftSeconds` – Number (double), not more than None.

  The estimated time left to complete a running crawl.
- `StillEstimating` – Boolean.

  True if the crawler is still estimating how long it will take to complete this run.
- `LastRuntimeSeconds` – Number (double), not more than None.

  The duration of the crawler's most recent run, in seconds.
- `MedianRuntimeSeconds` – Number (double), not more than None.

  The median duration of this crawler's runs, in seconds.
- `TablesCreated` – Number (integer), not more than None.

  The number of tables created by this crawler.
- `TablesUpdated` – Number (integer), not more than None.

  The number of tables updated by this crawler.
- `TablesDeleted` – Number (integer), not more than None.

  The number of tables deleted by this crawler.

CrawlerHistory structure

Contains the information for a run of a crawler.

Fields

- `CrawlId` – UTF-8 string.

A UUID identifier for each crawl.

- `State` – UTF-8 string (valid values: `RUNNING` | `COMPLETED` | `FAILED` | `STOPPED`).

  The state of the crawl.
- `StartTime` – Timestamp.

  The date and time on which the crawl started.
- `EndTime` – Timestamp.

  The date and time on which the crawl ended.
- `Summary` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A run summary for the specific crawl in JSON. Contains the catalog tables and partitions that were added, updated, or deleted.
- `ErrorMessage` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  If an error occurred, the error message associated with the crawl.
- `LogGroup` – UTF-8 string, not less than 1 or more than 512 bytes long, matching the Log group string pattern.

  The log group associated with the crawl.
- `LogStream` – UTF-8 string, not less than 1 or more than 512 bytes long, matching the Log-stream string pattern.

  The log stream associated with the crawl.
- `MessagePrefix` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The prefix for a CloudWatch message about this crawl.
- `DPUHour` – Number (double), not more than None.

  The number of data processing units (DPU) used in hours for the crawl.

CrawlsFilter structure

A list of fields, comparators and value that you can use to filter the crawler runs for a specified crawler.

Fields

- `FieldName` – UTF-8 string (valid values: `CRAWL_ID` | `STATE` | `START_TIME` | `END_TIME` | `DPU_HOUR`).

  A key used to filter the crawler runs for a specified crawler. Valid values for each of the field names are:

  - `CRAWL_ID`: A string representing the UUID identifier for a crawl.
  - `STATE`: A string representing the state of the crawl.
  - `START_TIME` and `END_TIME`: The epoch timestamp in milliseconds.
  - `DPU_HOUR`: The number of data processing unit (DPU) hours used for the crawl.
- `FilterOperator` – UTF-8 string (valid values: `GT` | `GE` | `LT` | `LE` | `EQ` | `NE`).

  A defined comparator that operates on the value. The available operators are:

  - `GT`: Greater than.
  - `GE`: Greater than or equal to.
  - `LT`: Less than.
  - `LE`: Less than or equal to.
  - `EQ`: Equal to.
  - `NE`: Not equal to.
- `FieldValue` – UTF-8 string.

  The value provided for comparison on the crawl field.

SchemaChangePolicy structure

A policy that specifies update and deletion behaviors for the crawler.

Fields

- `UpdateBehavior` – UTF-8 string (valid values: `LOG` | `UPDATE_IN_DATABASE`).

The update behavior when the crawler finds a changed schema.

- `DeleteBehavior` – UTF-8 string (valid values: `LOG` | `DELETE_FROM_DATABASE` | `DEPRECATE_IN_DATABASE`).

  The deletion behavior when the crawler finds a deleted object.

LastCrawlInfo structure

Status and error information about the most recent crawl.

Fields

- `Status` – UTF-8 string (valid values: `SUCCEEDED` | `CANCELLED` | `FAILED`).

  Status of the last crawl.
- `ErrorMessage` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  If an error occurred, the error information about the last crawl.
- `LogGroup` – UTF-8 string, not less than 1 or more than 512 bytes long, matching the Log group string pattern.

  The log group for the last crawl.
- `LogStream` – UTF-8 string, not less than 1 or more than 512 bytes long, matching the Log-stream string pattern.

  The log stream for the last crawl.
- `MessagePrefix` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The prefix for a message about this crawl.
- `StartTime` – Timestamp.

  The time at which the crawl started.

RecrawlPolicy structure

When crawling an Amazon S3 data source after the first crawl is complete, specifies whether to crawl the entire dataset again or to crawl only folders that were added since the last crawler run. For more information, see Incremental Crawls in AWS Glue in the developer guide.

Fields

- `RecrawlBehavior` – UTF-8 string (valid values: `CRAWL_EVERYTHING` | `CRAWL_NEW_FOLDERS_ONLY` | `CRAWL_EVENT_MODE`).

  Specifies whether to crawl the entire dataset again or to crawl only folders that were added since the last crawler run.

  A value of `CRAWL_EVERYTHING` specifies crawling the entire dataset again.

  A value of `CRAWL_NEW_FOLDERS_ONLY` specifies crawling only folders that were added since the last crawler run.

  A value of `CRAWL_EVENT_MODE` specifies crawling only the changes identified by Amazon S3 events.

LineageConfiguration structure

Specifies data lineage configuration settings for the crawler.

Fields

- `CrawlerLineageSettings` – UTF-8 string (valid values: `ENABLE` | `DISABLE`).

  Specifies whether data lineage is enabled for the crawler. Valid values are:

  - ENABLE: enables data lineage for the crawler
  - DISABLE: disables data lineage for the crawler

LakeFormationConfiguration structure

Specifies AWS Lake Formation configuration settings for the crawler.

Fields

- `UseLakeFormationCredentials` – Boolean.

  Specifies whether to use AWS Lake Formation credentials for the crawler instead of the IAM role credentials.
- `AccountId` – UTF-8 string, not more than 12 bytes long.

  Required for cross account crawls. For same account crawls as the target data, this can be left as null.

Operations

- CreateCrawler action (Python: create_crawler)
- DeleteCrawler action (Python: delete_crawler)
- GetCrawler action (Python: get_crawler)
- GetCrawlers action (Python: get_crawlers)
- GetCrawlerMetrics action (Python: get_crawler_metrics)
- UpdateCrawler action (Python: update_crawler)
- StartCrawler action (Python: start_crawler)
- StopCrawler action (Python: stop_crawler)
- BatchGetCrawlers action (Python: batch_get_crawlers)
- ListCrawlers action (Python: list_crawlers)
- ListCrawls action (Python: list_crawls)

CreateCrawler action (Python: create_crawler)

Creates a new crawler with specified targets, role, configuration, and optional schedule. At least one crawl target must be specified, in the `s3Targets` field, the `jdbcTargets` field, or the `DynamoDBTargets` field.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

Name of the new crawler.

- `Role` – *Required:* UTF-8 string.

  The IAM role or Amazon Resource Name (ARN) of an IAM role used by the new crawler to access customer resources.
- `DatabaseName` – UTF-8 string.

  The AWS Glue database where results are written, such as: `arn:aws:daylight:us-east-1::database/sometable/*`.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the new crawler.
- `Targets` – *Required:* A CrawlerTargets object.

  A list of collection of targets to crawl.
- `Schedule` – UTF-8 string.

  A `cron` expression used to specify the schedule (see Time-Based Schedules for Jobs and Crawlers. For example, to run something every day at 12:15 UTC, you would specify: `cron(15 12 * * ? *)`.
- `Classifiers` – An array of UTF-8 strings.

  A list of custom classifiers that the user has registered. By default, all built-in classifiers are included in a crawl, but these custom classifiers always override the default classifiers for a given classification.
- `TablePrefix` – UTF-8 string, not more than 128 bytes long.

  The table prefix used for catalog tables that are created.
- `SchemaChangePolicy` – A SchemaChangePolicy object.

  The policy for the crawler's update and deletion behavior.
- `RecrawlPolicy` – A RecrawlPolicy object.

  A policy that specifies whether to crawl the entire dataset again, or to crawl only folders that were added since the last crawler run.
- `LineageConfiguration` – A LineageConfiguration object.

Specifies data lineage configuration settings for the crawler.

- `LakeFormationConfiguration` – A LakeFormationConfiguration object.

  Specifies AWS Lake Formation configuration settings for the crawler.
- `Configuration` – UTF-8 string.

  Crawler configuration information. This versioned JSON string allows users to specify aspects of a crawler's behavior. For more information, see Setting crawler configuration options.
- `CrawlerSecurityConfiguration` – UTF-8 string, not more than 128 bytes long.

  The name of the `SecurityConfiguration` structure to be used by this crawler.
- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  The tags to use with this crawler request. You may use tags to limit access to the crawler. For more information about tags in AWS Glue, see AWS Tags in AWS Glue in the developer guide.

Response

- *No Response parameters.*

Errors

- `InvalidInputException`
- `AlreadyExistsException`
- `OperationTimeoutException`
- `ResourceNumberLimitExceededException`

DeleteCrawler action (Python: delete_crawler)

Removes a specified crawler from the AWS Glue Data Catalog, unless the crawler state is `RUNNING`.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the crawler to remove.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `CrawlerRunningException`
- `SchedulerTransitioningException`
- `OperationTimeoutException`

GetCrawler action (Python: get_crawler)

Retrieves metadata for a specified crawler.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the crawler to retrieve metadata for.

Response

- `Crawler` – A Crawler object.

  The metadata for the specified crawler.

Errors

- `EntityNotFoundException`

- `OperationTimeoutException`

GetCrawlers action (Python: get_crawlers)

Retrieves metadata for all crawlers defined in the customer account.

Request

- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The number of crawlers to return on each call.
- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation request.

Response

- `Crawlers` – An array of [Crawler](#) objects.

  A list of crawler metadata.
- `NextToken` – UTF-8 string.

  A continuation token, if the returned list has not reached the end of those defined in this customer account.

Errors

- `OperationTimeoutException`

GetCrawlerMetrics action (Python: get_crawler_metrics)

Retrieves metrics about specified crawlers.

Request

- `CrawlerNameList` – An array of UTF-8 strings, not more than 100 strings.

  A list of the names of crawlers about which to retrieve metrics.
- `MaxResults` – Number (integer), not less than 1 or more than 1000.

The maximum size of a list to return.

- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation call.

Response

- `CrawlerMetricsList` – An array of CrawlerMetrics objects.

  A list of metrics for the specified crawler.
- `NextToken` – UTF-8 string.

  A continuation token, if the returned list does not contain the last metric available.

Errors

- `OperationTimeoutException`

UpdateCrawler action (Python: update_crawler)

Updates a crawler. If a crawler is running, you must stop it using `StopCrawler` before updating it.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Name of the new crawler.
- `Role` – UTF-8 string.

  The IAM role or Amazon Resource Name (ARN) of an IAM role that is used by the new crawler to access customer resources.
- `DatabaseName` – UTF-8 string.

  The AWS Glue database where results are stored, such as: `arn:aws:daylight:us-east-1::database/sometable/*`.

- `Description` – UTF-8 string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the new crawler.
- `Targets` – A CrawlerTargets object.

  A list of targets to crawl.
- `Schedule` – UTF-8 string.

  A `cron` expression used to specify the schedule (see Time-Based Schedules for Jobs and Crawlers. For example, to run something every day at 12:15 UTC, you would specify: `cron(15 12 * * ? *)`.
- `Classifiers` – An array of UTF-8 strings.

  A list of custom classifiers that the user has registered. By default, all built-in classifiers are included in a crawl, but these custom classifiers always override the default classifiers for a given classification.
- `TablePrefix` – UTF-8 string, not more than 128 bytes long.

  The table prefix used for catalog tables that are created.
- `SchemaChangePolicy` – A SchemaChangePolicy object.

  The policy for the crawler's update and deletion behavior.
- `RecrawlPolicy` – A RecrawlPolicy object.

  A policy that specifies whether to crawl the entire dataset again, or to crawl only folders that were added since the last crawler run.
- `LineageConfiguration` – A LineageConfiguration object.

  Specifies data lineage configuration settings for the crawler.
- `LakeFormationConfiguration` – A LakeFormationConfiguration object.

  Specifies AWS Lake Formation configuration settings for the crawler.
- `Configuration` – UTF-8 string.

  Crawler configuration information. This versioned JSON string allows users to specify aspects of a crawler's behavior. For more information, see Setting crawler configuration options.
- `CrawlerSecurityConfiguration` – UTF-8 string, not more than 128 bytes long.

The name of the `SecurityConfiguration` structure to be used by this crawler.

Response

- *No Response parameters.*

Errors

- `InvalidInputException`
- `VersionMismatchException`
- `EntityNotFoundException`
- `CrawlerRunningException`
- `OperationTimeoutException`

StartCrawler action (Python: start_crawler)

Starts a crawl using the specified crawler, regardless of what is scheduled. If the crawler is already running, returns a CrawlerRunningException.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Name of the crawler to start.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `CrawlerRunningException`
- `OperationTimeoutException`

StopCrawler action (Python: stop_crawler)

If the specified crawler is running, stops the crawl.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Name of the crawler to stop.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `CrawlerNotRunningException`
- `CrawlerStoppingException`
- `OperationTimeoutException`

BatchGetCrawlers action (Python: batch_get_crawlers)

Returns a list of resource metadata for a given list of crawler names. After calling the `ListCrawlers` operation, you can call this operation to access the data to which you have been granted permissions. This operation supports all IAM permissions, including permission conditions that uses tags.

Request

- `CrawlerNames` – *Required:* An array of UTF-8 strings, not more than 100 strings.

  A list of crawler names, which might be the names returned from the `ListCrawlers` operation.

Response

- `Crawlers` – An array of Crawler objects.

A list of crawler definitions.

- `CrawlersNotFound` – An array of UTF-8 strings, not more than 100 strings.

  A list of names of crawlers that were not found.

Errors

- `InvalidInputException`
- `OperationTimeoutException`

ListCrawlers action (Python: list_crawlers)

Retrieves the names of all crawler resources in this AWS account, or the resources with the specified tag. This operation allows you to see which resources are available in your account, and their names.

This operation takes the optional `Tags` field, which you can use as a filter on the response so that tagged resources can be retrieved as a group. If you choose to use tags filtering, only resources with the tag are retrieved.

Request

- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum size of a list to return.
- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation request.
- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  Specifies to return only these tagged resources.

Response

- `CrawlerNames` – An array of UTF-8 strings, not more than 100 strings.

  The names of all crawlers in the account, or the crawlers with the specified tags.
- `NextToken` – UTF-8 string.

  A continuation token, if the returned list does not contain the last metric available.

Errors

- `OperationTimeoutException`

ListCrawls action (Python: list_crawls)

Returns all the crawls of a specified crawler. Returns only the crawls that have occurred since the launch date of the crawler history feature, and only retains up to 12 months of crawls. Older crawls will not be returned.

You may use this API to:

- Retrive all the crawls of a specified crawler.
- Retrieve all the crawls of a specified crawler within a limited count.
- Retrieve all the crawls of a specified crawler in a specific time range.
- Retrieve all the crawls of a specified crawler with a particular state, crawl ID, or DPU hour value.

Request

- `CrawlerName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the crawler whose runs you want to retrieve.
- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum number of results to return. The default is 20, and maximum is 100.
- `Filters` – An array of CrawlsFilter objects.

  Filters the crawls by the criteria you specify in a list of `CrawlsFilter` objects.

- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation call.

Response

- `Crawls` – An array of [CrawlerHistory](#) objects.

  A list of `CrawlerHistory` objects representing the crawl runs that meet your criteria.
- `NextToken` – UTF-8 string.

  A continuation token for paginating the returned list of tokens, returned if the current segment of the list is not the last.

Errors

- `EntityNotFoundException`
- `OperationTimeoutException`
- `InvalidInputException`

Column statistics API

The column statistics API describes AWS Glue APIs for returning statistics on columns in a table.

Data types

- [ColumnStatisticsTaskRun structure](#)
- [ColumnStatisticsTaskSettings structure](#)
- [ExecutionAttempt structure](#)

ColumnStatisticsTaskRun structure

The object that shows the details of the column stats run.

Fields

- `CustomerId` – UTF-8 string, not more than 12 bytes long.

The AWS account ID.

- `ColumnStatisticsTaskRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The identifier for the particular column statistics task run.
- `DatabaseName` – UTF-8 string.

  The database where the table resides.
- `TableName` – UTF-8 string.

  The name of the table for which column statistics is generated.
- `ColumnNameList` – An array of UTF-8 strings.

  A list of the column names. If none is supplied, all column names for the table will be used by default.
- `CatalogID` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog where the table resides. If none is supplied, the AWS account ID is used by default.
- `Role` – UTF-8 string.

  The IAM role that the service assumes to generate statistics.
- `SampleSize` – Number (double), not more than 100.

  The percentage of rows used to generate statistics. If none is supplied, the entire table will be used to generate stats.
- `SecurityConfiguration` – UTF-8 string, not more than 128 bytes long.

  Name of the security configuration that is used to encrypt CloudWatch logs for the column stats task run.
- `NumberOfWorkers` – Number (integer), at least 1.

  The number of workers used to generate column statistics. The job is preconfigured to autoscale up to 25 instances.
- `WorkerType` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The type of workers being used for generating stats. The default is `g.1x`.

- `ComputationType` – UTF-8 string (valid values: `FULL` | `INCREMENTAL`).

  The type of column statistics computation.
- `Status` – UTF-8 string (valid values: `STARTING` | `RUNNING` | `SUCCEEDED` | `FAILED` | `STOPPED`).

  The status of the task run.
- `CreationTime` – Timestamp.

  The time that this task was created.
- `LastUpdated` – Timestamp.

  The last point in time when this task was modified.
- `StartTime` – Timestamp.

  The start time of the task.
- `EndTime` – Timestamp.

  The end time of the task.
- `ErrorMessage` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  The error message for the job.
- `DPUSeconds` – Number (double), not more than None.

  The calculated DPU usage in seconds for all autoscaled workers.

ColumnStatisticsTaskSettings structure

The settings for a column statistics task.

Fields

- `DatabaseName` – UTF-8 string.

  The name of the database where the table resides.
- `TableName` – UTF-8 string.

The name of the table for which to generate column statistics.

- `Schedule` – A Schedule object.

  A schedule for running the column statistics, specified in CRON syntax.
- `ColumnNameList` – An array of UTF-8 strings.

  A list of column names for which to run statistics.
- `CatalogID` – Catalog id string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog in which the database resides.
- `Role` – UTF-8 string.

  The role used for running the column statistics.
- `SampleSize` – Number (double), not more than 100.

  The percentage of data to sample.
- `SecurityConfiguration` – UTF-8 string, not more than 128 bytes long.

  Name of the security configuration that is used to encrypt CloudWatch logs.
- `ScheduleType` – UTF-8 string (valid values: `CRON` | `AUTO`).

  The type of schedule for a column statistics task. Possible values may be `CRON` or `AUTO`.
- `SettingSource` – UTF-8 string (valid values: `CATALOG` | `TABLE`).

  The source of setting the column statistics task. Possible values may be `CATALOG` or `TABLE`.
- `LastExecutionAttempt` – An ExecutionAttempt object.

  The last `ExecutionAttempt` for the column statistics task run.

ExecutionAttempt structure

A run attempt for a column statistics task run.

Fields

- `Status` – UTF-8 string (valid values: `FAILED` | `STARTED`).

The status of the last column statistics task run.

- `ColumnStatisticsTaskRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A task run ID for the last column statistics task run.

- `ExecutionTimestamp` – Timestamp.

  A timestamp when the last column statistics task run occurred.

- `ErrorMessage` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  An error message associated with the last column statistics task run.

Operations

- StartColumnStatisticsTaskRun action (Python: start_column_statistics_task_run)
- GetColumnStatisticsTaskRun action (Python: get_column_statistics_task_run)
- GetColumnStatisticsTaskRuns action (Python: get_column_statistics_task_runs)
- ListColumnStatisticsTaskRuns action (Python: list_column_statistics_task_runs)
- StopColumnStatisticsTaskRun action (Python: stop_column_statistics_task_run)
- CreateColumnStatisticsTaskSettings action (Python: create_column_statistics_task_settings)
- UpdateColumnStatisticsTaskSettings action (Python: update_column_statistics_task_settings)
- GetColumnStatisticsTaskSettings action (Python: get_column_statistics_task_settings)
- DeleteColumnStatisticsTaskSettings action (Python: delete_column_statistics_task_settings)
- StartColumnStatisticsTaskRunSchedule action (Python: start_column_statistics_task_run_schedule)
- StopColumnStatisticsTaskRunSchedule action (Python: stop_column_statistics_task_run_schedule)

StartColumnStatisticsTaskRun action (Python: start_column_statistics_task_run)

Starts a column statistics task run, for a specified table and columns.

Request

- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the database where the table resides.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table to generate statistics.
- `ColumnNameList` – An array of UTF-8 strings.

  A list of the column names to generate statistics. If none is supplied, all column names for the table will be used by default.
- `Role` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The IAM role that the service assumes to generate statistics.
- `SampleSize` – Number (double), not more than 100.

  The percentage of rows used to generate statistics. If none is supplied, the entire table will be used to generate stats.
- `CatalogID` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog where the table reside. If none is supplied, the AWS account ID is used by default.
- `SecurityConfiguration` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Name of the security configuration that is used to encrypt CloudWatch logs for the column stats task run.

Response

- `ColumnStatisticsTaskRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The identifier for the column statistics task run.

Errors

- `AccessDeniedException`
- `EntityNotFoundException`
- `ColumnStatisticsTaskRunningException`
- `OperationTimeoutException`
- `ResourceNumberLimitExceededException`
- `InvalidInputException`

GetColumnStatisticsTaskRun action (Python: get_column_statistics_task_run)

Get the associated metadata/information for a task run, given a task run ID.

Request

- `ColumnStatisticsTaskRunId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The identifier for the particular column statistics task run.

Response

- `ColumnStatisticsTaskRun` – A ColumnStatisticsTaskRun object.

  A `ColumnStatisticsTaskRun` object representing the details of the column stats run.

Errors

- `EntityNotFoundException`
- `OperationTimeoutException`
- `InvalidInputException`

GetColumnStatisticsTaskRuns action (Python: get_column_statistics_task_runs)

Retrieves information about all runs associated with the specified table.

Request

- `DatabaseName` – *Required:* UTF-8 string.

  The name of the database where the table resides.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table.
- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum size of the response.
- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation call.

Response

- `ColumnStatisticsTaskRuns` – An array of ColumnStatisticsTaskRun objects.

  A list of column statistics task runs.
- `NextToken` – UTF-8 string.

  A continuation token, if not all task runs have yet been returned.

Errors

- `OperationTimeoutException`

ListColumnStatisticsTaskRuns action (Python: list_column_statistics_task_runs)

List all task runs for a particular account.

Request

- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum size of the response.
- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation call.

Response

- `ColumnStatisticsTaskRunIds` – An array of UTF-8 strings, not more than 100 strings.

  A list of column statistics task run IDs.
- `NextToken` – UTF-8 string.

  A continuation token, if not all task run IDs have yet been returned.

Errors

- `OperationTimeoutException`

StopColumnStatisticsTaskRun action (Python: stop_column_statistics_task_run)

Stops a task run for the specified table.

Request

- `DatabaseName` – *Required:* UTF-8 string.

  The name of the database where the table resides.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `ColumnStatisticsTaskNotRunningException`
- `ColumnStatisticsTaskStoppingException`
- `OperationTimeoutException`

CreateColumnStatisticsTaskSettings action (Python: create_column_statistics_task_settings)

Creates settings for a column statistics task.

Request

- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the database where the table resides.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table for which to generate column statistics.
- `Role` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The role used for running the column statistics.
- `Schedule` – UTF-8 string.

  A schedule for running the column statistics, specified in CRON syntax.
- `ColumnNameList` – An array of UTF-8 strings.

  A list of column names for which to run statistics.
- `SampleSize` – Number (double), not more than 100.

  The percentage of data to sample.
- `CatalogID` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The ID of the Data Catalog in which the database resides.

- `SecurityConfiguration` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Name of the security configuration that is used to encrypt CloudWatch logs.
- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  A map of tags.

Response

- *No Response parameters.*

Errors

- `AlreadyExistsException`
- `AccessDeniedException`
- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`
- `ResourceNumberLimitExceededException`
- `ColumnStatisticsTaskRunningException`

UpdateColumnStatisticsTaskSettings action (Python: update_column_statistics_task_settings)

Updates settings for a column statistics task.

Request

- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The name of the database where the table resides.

- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table for which to generate column statistics.
- `Role` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The role used for running the column statistics.
- `Schedule` – UTF-8 string.

  A schedule for running the column statistics, specified in CRON syntax.
- `ColumnNameList` – An array of UTF-8 strings.

  A list of column names for which to run statistics.
- `SampleSize` – Number (double), not more than 100.

  The percentage of data to sample.
- `CatalogID` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the Data Catalog in which the database resides.
- `SecurityConfiguration` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Name of the security configuration that is used to encrypt CloudWatch logs.

Response

- *No Response parameters.*

Errors

- `AccessDeniedException`
- `EntityNotFoundException`
- `InvalidInputException`

- `VersionMismatchException`
- `OperationTimeoutException`

GetColumnStatisticsTaskSettings action (Python: get_column_statistics_task_settings)

Gets settings for a column statistics task.

Request

- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the database where the table resides.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table for which to retrieve column statistics.

Response

- `ColumnStatisticsTaskSettings` – A ColumnStatisticsTaskSettings object.

  A `ColumnStatisticsTaskSettings` object representing the settings for the column statistics task.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`

DeleteColumnStatisticsTaskSettings action (Python: delete_column_statistics_task_settings)

Deletes settings for a column statistics task.

Request

- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the database where the table resides.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table for which to delete column statistics.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`

StartColumnStatisticsTaskRunSchedule action (Python: start_column_statistics_task_run_schedule)

Starts a column statistics task run schedule.

Request

- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the database where the table resides.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table for which to start a column statistic task run schedule.

Response

- *No Response parameters.*

Errors

- `AccessDeniedException`
- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`

StopColumnStatisticsTaskRunSchedule action (Python: stop_column_statistics_task_run_schedule)

Stops a column statistics task run schedule.

Request

- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the database where the table resides.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table for which to stop a column statistic task run schedule.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`

Exceptions

- ColumnStatisticsTaskRunningException structure
- ColumnStatisticsTaskNotRunningException structure
- ColumnStatisticsTaskStoppingException structure
- ColumnStatisticsTaskAutoConcurrencyLimitException structure
- InvalidCatalogSettingException structure

ColumnStatisticsTaskRunningException structure

An exception thrown when you try to start another job while running a column stats generation job.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

ColumnStatisticsTaskNotRunningException structure

An exception thrown when you try to stop a task run when there is no task running.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

ColumnStatisticsTaskStoppingException structure

An exception thrown when you try to stop a task run.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

ColumnStatisticsTaskAutoConcurrencyLimitException structure

An exception thrown when you have already reached the limit of concurrent auto statistics jobs.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

InvalidCatalogSettingException structure

An exception thrown when there is a problem with the catalog settings.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

Crawler scheduler API

The Crawler scheduler API describes AWS Glue crawler data types, along with the API for creating, deleting, updating, and listing crawlers.

Data types

- Schedule structure

Schedule structure

A scheduling object using a `cron` statement to schedule an event.

Fields

- `ScheduleExpression` – UTF-8 string.

  A `cron` expression used to specify the schedule (see Time-Based Schedules for Jobs and Crawlers. For example, to run something every day at 12:15 UTC, you would specify: `cron(15 12 * * ? *)`.

- `State` – UTF-8 string (valid values: `SCHEDULED` | `NOT_SCHEDULED` | `TRANSITIONING`).

  The state of the schedule.

Operations

- UpdateCrawlerSchedule action (Python: update_crawler_schedule)
- StartCrawlerSchedule action (Python: start_crawler_schedule)
- StopCrawlerSchedule action (Python: stop_crawler_schedule)

UpdateCrawlerSchedule action (Python: update_crawler_schedule)

Updates the schedule of a crawler using a `cron` expression.

Request

- `CrawlerName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the crawler whose schedule to update.
- `Schedule` – UTF-8 string.

  The updated `cron` expression used to specify the schedule (see Time-Based Schedules for Jobs and Crawlers. For example, to run something every day at 12:15 UTC, you would specify: `cron(15 12 * * ? *)`.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `VersionMismatchException`
- `SchedulerTransitioningException`

- `OperationTimeoutException`

StartCrawlerSchedule action (Python: start_crawler_schedule)

Changes the schedule state of the specified crawler to `SCHEDULED`, unless the crawler is already running or the schedule state is already `SCHEDULED`.

Request

- `CrawlerName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Name of the crawler to schedule.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `SchedulerRunningException`
- `SchedulerTransitioningException`
- `NoScheduleException`
- `OperationTimeoutException`

StopCrawlerSchedule action (Python: stop_crawler_schedule)

Sets the schedule state of the specified crawler to `NOT_SCHEDULED`, but does not stop the crawler if it is already running.

Request

- `CrawlerName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Name of the crawler whose schedule state to set.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `SchedulerNotRunningException`
- `SchedulerTransitioningException`
- `OperationTimeoutException`

Autogenerating ETL Scripts API

The ETL script-generation API describes the datatypes and API for generating ETL scripts in AWS Glue.

Data types

- CodeGenNode structure
- CodeGenNodeArg structure
- CodeGenEdge structure
- Location structure
- CatalogEntry structure
- MappingEntry structure

CodeGenNode structure

Represents a node in a directed acyclic graph (DAG)

Fields

- `Id` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Identifier string pattern.

  A node identifier that is unique within the node's graph.
- `NodeType` – *Required:* UTF-8 string.

The type of node that this is.

- `Args` – *Required:* An array of [CodeGenNodeArg](#) objects, not more than 50 structures.

  Properties of the node, in the form of name-value pairs.
- `LineNumber` – Number (integer).

  The line number of the node.

CodeGenNodeArg structure

An argument or property of a node.

Fields

- `Name` – *Required:* UTF-8 string.

  The name of the argument or property.
- `Value` – *Required:* UTF-8 string.

  The value of the argument or property.
- `Param` – Boolean.

  True if the value is used as a parameter.

CodeGenEdge structure

Represents a directional edge in a directed acyclic graph (DAG).

Fields

- `Source` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Identifier string pattern.

  The ID of the node at which the edge starts.
- `Target` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Identifier string pattern.

  The ID of the node at which the edge ends.

- `TargetParameter` – UTF-8 string.

  The target of the edge.

Location structure

The location of resources.

Fields

- `Jdbc` – An array of CodeGenNodeArg objects, not more than 50 structures.

  A JDBC location.
- `S3` – An array of CodeGenNodeArg objects, not more than 50 structures.

  An Amazon Simple Storage Service (Amazon S3) location.
- `DynamoDB` – An array of CodeGenNodeArg objects, not more than 50 structures.

  An Amazon DynamoDB table location.

CatalogEntry structure

Specifies a table definition in the AWS Glue Data Catalog.

Fields

- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The database in which the table metadata resides.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the table in question.

MappingEntry structure

Defines a mapping.

Fields

- `SourceTable` – UTF-8 string.

  The name of the source table.
- `SourcePath` – UTF-8 string.

  The source path.
- `SourceType` – UTF-8 string.

  The source type.
- `TargetTable` – UTF-8 string.

  The target table.
- `TargetPath` – UTF-8 string.

  The target path.
- `TargetType` – UTF-8 string.

  The target type.

Operations

- CreateScript action (Python: create_script)
- GetDataflowGraph action (Python: get_dataflow_graph)
- GetMapping action (Python: get_mapping)
- GetPlan action (Python: get_plan)

CreateScript action (Python: create_script)

Transforms a directed acyclic graph (DAG) into code.

Request

- `DagNodes` – An array of CodeGenNode objects.

  A list of the nodes in the DAG.

- `DagEdges` – An array of CodeGenEdge objects.

  A list of the edges in the DAG.
- `Language` – UTF-8 string (valid values: `PYTHON` | `SCALA`).

  The programming language of the resulting code from the DAG.

Response

- `PythonScript` – UTF-8 string.

  The Python script generated from the DAG.
- `ScalaCode` – UTF-8 string.

  The Scala code generated from the DAG.

Errors

- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`

GetDataflowGraph action (Python: get_dataflow_graph)

Transforms a Python script into a directed acyclic graph (DAG).

Request

- `PythonScript` – UTF-8 string.

  The Python script to transform.

Response

- `DagNodes` – An array of CodeGenNode objects.

  A list of the nodes in the resulting DAG.
- `DagEdges` – An array of CodeGenEdge objects.

A list of the edges in the resulting DAG.

Errors

- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`

GetMapping action (Python: get_mapping)

Creates mappings.

Request

- `Source` – *Required:* A CatalogEntry object.

  Specifies the source table.
- `Sinks` – An array of CatalogEntry objects.

  A list of target tables.
- `Location` – A Location object.

  Parameters for the mapping.

Response

- `Mapping` – *Required:* An array of MappingEntry objects.

  A list of mappings to the specified targets.

Errors

- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `EntityNotFoundException`

GetPlan action (Python: get_plan)

Gets code to perform a specified mapping.

Request

- `Mapping` – *Required:* An array of MappingEntry objects.

  The list of mappings from a source table to target tables.
- `Source` – *Required:* A CatalogEntry object.

  The source table.
- `Sinks` – An array of CatalogEntry objects.

  The target tables.
- `Location` – A Location object.

  The parameters for the mapping.
- `Language` – UTF-8 string (valid values: `PYTHON` | `SCALA`).

  The programming language of the code to perform the mapping.
- `AdditionalPlanOptionsMap` – A map array of key-value pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

  A map to hold additional optional key-value parameters.

  Currently, these key-value pairs are supported:
  - `inferSchema` — Specifies whether to set `inferSchema` to true or false for the default script generated by an AWS Glue job. For example, to set `inferSchema` to true, pass the following key value pair:

    ```
    --additional-plan-options-map '{"inferSchema":"true"}'
    ```

Response

- `PythonScript` – UTF-8 string.

A Python script to perform the mapping.

- `ScalaCode` – UTF-8 string.

  The Scala code to perform the mapping.

Errors

- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`

Visual job API

The Visual job API allows you to create data integration jobs by using the AWS Glue API from a JSON object that represents a visual configuration of a AWS Glue job.

A list of `CodeGenConfigurationNodes` are provided to a create or update job API to register a DAG in AWS Glue Studio for the created job and generate the associated code.

Data types

- CodeGenConfigurationNode structure
- JDBCConnectorOptions structure
- StreamingDataPreviewOptions structure
- AthenaConnectorSource structure
- JDBCConnectorSource structure
- SparkConnectorSource structure
- CatalogSource structure
- MySQLCatalogSource structure
- PostgreSQLCatalogSource structure
- OracleSQLCatalogSource structure
- MicrosoftSQLServerCatalogSource structure
- CatalogKinesisSource structure

- DirectKinesisSource structure
- KinesisStreamingSourceOptions structure
- CatalogKafkaSource structure
- DirectKafkaSource structure
- KafkaStreamingSourceOptions structure
- RedshiftSource structure
- AmazonRedshiftSource structure
- AmazonRedshiftNodeData structure
- AmazonRedshiftAdvancedOption structure
- Option structure
- S3CatalogSource structure
- S3SourceAdditionalOptions structure
- S3CsvSource structure
- DirectJDBCSource structure
- S3DirectSourceAdditionalOptions structure
- S3JsonSource structure
- S3ParquetSource structure
- S3DeltaSource structure
- S3CatalogDeltaSource structure
- CatalogDeltaSource structure
- S3HudiSource structure
- S3CatalogHudiSource structure
- S3ExcelSource structure
- CatalogHudiSource structure
- DynamoDBCatalogSource structure
- RelationalCatalogSource structure
- JDBCConnectorTarget structure
- SparkConnectorTarget structure
- BasicCatalogTarget structure
- MySQLCatalogTarget structure

- PostgreSQLCatalogTarget structure
- OracleSQLCatalogTarget structure
- MicrosoftSQLServerCatalogTarget structure
- RedshiftTarget structure
- AmazonRedshiftTarget structure
- UpsertRedshiftTargetOptions structure
- S3CatalogTarget structure
- S3GlueParquetTarget structure
- CatalogSchemaChangePolicy structure
- S3DirectTarget structure
- S3HudiCatalogTarget structure
- S3HudiDirectTarget structure
- S3DeltaCatalogTarget structure
- S3DeltaDirectTarget structure
- S3HyperDirectTarget structure
- S3IcebergDirectTarget structure
- DirectSchemaChangePolicy structure
- ApplyMapping structure
- Mapping structure
- SelectFields structure
- DropFields structure
- RenameField structure
- Spigot structure
- Join structure
- JoinColumn structure
- SplitFields structure
- SelectFromCollection structure
- FillMissingValues structure
- Filter structure
- FilterExpression structure

- FilterValue structure
- CustomCode structure
- SparkSQL structure
- SqlAlias structure
- DropNullFields structure
- NullCheckBoxList structure
- NullValueField structure
- Datatype structure
- Merge structure
- Union structure
- PIIDetection structure
- Aggregate structure
- DropDuplicates structure
- GovernedCatalogTarget structure
- GovernedCatalogSource structure
- AggregateOperation structure
- GlueSchema structure
- GlueStudioSchemaColumn structure
- GlueStudioColumn structure
- DynamicTransform structure
- TransformConfigParameter structure
- EvaluateDataQuality structure
- DQResultsPublishingOptions structure
- DQStopJobOnFailureOptions structure
- EvaluateDataQualityMultiFrame structure
- Recipe structure
- RecipeReference structure
- SnowflakeNodeData structure
- SnowflakeSource structure
- SnowflakeTarget structure

- ConnectorDataSource structure
- ConnectorDataTarget structure
- RecipeStep structure
- RecipeAction structure
- ConditionExpression structure
- S3CatalogIcebergSource structure
- CatalogIcebergSource structure
- S3IcebergCatalogTarget structure
- DynamoDBELTConnectorSource structure
- DDBELTConnectionOptions structure
- DDBELTCatalogAdditionalOptions structure
- Route structure
- GroupFilters structure
- AutoDataQuality structure

CodeGenConfigurationNode structure

`CodeGenConfigurationNode` enumerates all valid Node types. One and only one of its member variables can be populated.

Fields

- `AthenaConnectorSource` – An AthenaConnectorSource object.

  Specifies a connector to an Amazon Athena data source.
- `JDBCConnectorSource` – A JDBCConnectorSource object.

  Specifies a connector to a JDBC data source.
- `SparkConnectorSource` – A SparkConnectorSource object.

  Specifies a connector to an Apache Spark data source.
- `CatalogSource` – A CatalogSource object.

  Specifies a data store in the AWS Glue Data Catalog.
- `RedshiftSource` – A RedshiftSource object.

Specifies an Amazon Redshift data store.

- `S3CatalogSource` – A S3CatalogSource object.

  Specifies an Amazon S3 data store in the AWS Glue Data Catalog.
- `S3CsvSource` – A S3CsvSource object.

  Specifies a command-separated value (CSV) data store stored in Amazon S3.
- `S3JsonSource` – A S3JsonSource object.

  Specifies a JSON data store stored in Amazon S3.
- `S3ParquetSource` – A S3ParquetSource object.

  Specifies an Apache Parquet data store stored in Amazon S3.
- `RelationalCatalogSource` – A RelationalCatalogSource object.

  Specifies a relational catalog data store in the AWS Glue Data Catalog.
- `DynamoDBCatalogSource` – A DynamoDBCatalogSource object.

  Specifies a DynamoDBC Catalog data store in the AWS Glue Data Catalog.
- `JDBCConnectorTarget` – A JDBCConnectorTarget object.

  Specifies a data target that writes to Amazon S3 in Apache Parquet columnar storage.
- `SparkConnectorTarget` – A SparkConnectorTarget object.

  Specifies a target that uses an Apache Spark connector.
- `CatalogTarget` – A BasicCatalogTarget object.

  Specifies a target that uses a AWS Glue Data Catalog table.
- `RedshiftTarget` – A RedshiftTarget object.

  Specifies a target that uses Amazon Redshift.
- `S3CatalogTarget` – A S3CatalogTarget object.

  Specifies a data target that writes to Amazon S3 using the AWS Glue Data Catalog.
- `S3GlueParquetTarget` – A S3GlueParquetTarget object.

  Specifies a data target that writes to Amazon S3 in Apache Parquet columnar storage.

- `S3DirectTarget` – A S3DirectTarget object.

  Specifies a data target that writes to Amazon S3.
- `ApplyMapping` – An ApplyMapping object.

  Specifies a transform that maps data property keys in the data source to data property keys in the data target. You can rename keys, modify the data types for keys, and choose which keys to drop from the dataset.
- `SelectFields` – A SelectFields object.

  Specifies a transform that chooses the data property keys that you want to keep.
- `DropFields` – A DropFields object.

  Specifies a transform that chooses the data property keys that you want to drop.
- `RenameField` – A RenameField object.

  Specifies a transform that renames a single data property key.
- `Spigot` – A Spigot object.

  Specifies a transform that writes samples of the data to an Amazon S3 bucket.
- `Join` – A Join object.

  Specifies a transform that joins two datasets into one dataset using a comparison phrase on the specified data property keys. You can use inner, outer, left, right, left semi, and left anti joins.
- `SplitFields` – A SplitFields object.

  Specifies a transform that splits data property keys into two `DynamicFrames`. The output is a collection of `DynamicFrames`: one with selected data property keys, and one with the remaining data property keys.
- `SelectFromCollection` – A SelectFromCollection object.

  Specifies a transform that chooses one `DynamicFrame` from a collection of `DynamicFrames`. The output is the selected `DynamicFrame`
- `FillMissingValues` – A FillMissingValues object.

  Specifies a transform that locates records in the dataset that have missing values and adds a new field with a value determined by imputation. The input data set is used to train the machine learning model that determines what the missing value should be.

- `Filter` – A Filter object.

  Specifies a transform that splits a dataset into two, based on a filter condition.
- `CustomCode` – A CustomCode object.

  Specifies a transform that uses custom code you provide to perform the data transformation. The output is a collection of DynamicFrames.
- `SparkSQL` – A SparkSQL object.

  Specifies a transform where you enter a SQL query using Spark SQL syntax to transform the data. The output is a single `DynamicFrame`.
- `DirectKinesisSource` – A DirectKinesisSource object.

  Specifies a direct Amazon Kinesis data source.
- `DirectKafkaSource` – A DirectKafkaSource object.

  Specifies an Apache Kafka data store.
- `CatalogKinesisSource` – A CatalogKinesisSource object.

  Specifies a Kinesis data source in the AWS Glue Data Catalog.
- `CatalogKafkaSource` – A CatalogKafkaSource object.

  Specifies an Apache Kafka data store in the Data Catalog.
- `DropNullFields` – A DropNullFields object.

  Specifies a transform that removes columns from the dataset if all values in the column are 'null'. By default, AWS Glue Studio will recognize null objects, but some values such as empty strings, strings that are "null", -1 integers or other placeholders such as zeros, are not automatically recognized as nulls.
- `Merge` – A Merge object.

  Specifies a transform that merges a `DynamicFrame` with a staging `DynamicFrame` based on the specified primary keys to identify records. Duplicate records (records with the same primary keys) are not de-duplicated.
- `Union` – An Union object.

  Specifies a transform that combines the rows from two or more datasets into a single result.
- `PIIDetection` – A PIIDetection object.

Specifies a transform that identifies, removes or masks PII data.

- `Aggregate` – An Aggregate object.

  Specifies a transform that groups rows by chosen fields and computes the aggregated value by specified function.
- `DropDuplicates` – A DropDuplicates object.

  Specifies a transform that removes rows of repeating data from a data set.
- `GovernedCatalogTarget` – A GovernedCatalogTarget object.

  Specifies a data target that writes to a goverened catalog.
- `GovernedCatalogSource` – A GovernedCatalogSource object.

  Specifies a data source in a goverened Data Catalog.
- `MicrosoftSQLServerCatalogSource` – A MicrosoftSQLServerCatalogSource object.

  Specifies a Microsoft SQL server data source in the AWS Glue Data Catalog.
- `MySQLCatalogSource` – A MySQLCatalogSource object.

  Specifies a MySQL data source in the AWS Glue Data Catalog.
- `OracleSQLCatalogSource` – An OracleSQLCatalogSource object.

  Specifies an Oracle data source in the AWS Glue Data Catalog.
- `PostgreSQLCatalogSource` – A PostgreSQLCatalogSource object.

  Specifies a PostgresSQL data source in the AWS Glue Data Catalog.
- `MicrosoftSQLServerCatalogTarget` – A MicrosoftSQLServerCatalogTarget object.

  Specifies a target that uses Microsoft SQL.
- `MySQLCatalogTarget` – A MySQLCatalogTarget object.

  Specifies a target that uses MySQL.
- `OracleSQLCatalogTarget` – An OracleSQLCatalogTarget object.

  Specifies a target that uses Oracle SQL.
- `PostgreSQLCatalogTarget` – A PostgreSQLCatalogTarget object.

Specifies a target that uses Postgres SQL.

- `Route` – A Route object.

  Specifies a route node that directs data to different output paths based on defined filtering conditions.
- `DynamicTransform` – A DynamicTransform object.

  Specifies a custom visual transform created by a user.
- `EvaluateDataQuality` – An EvaluateDataQuality object.

  Specifies your data quality evaluation criteria.
- `S3CatalogHudiSource` – A S3CatalogHudiSource object.

  Specifies a Hudi data source that is registered in the AWS Glue Data Catalog. The data source must be stored in Amazon S3.
- `CatalogHudiSource` – A CatalogHudiSource object.

  Specifies a Hudi data source that is registered in the AWS Glue Data Catalog.
- `S3HudiSource` – A S3HudiSource object.

  Specifies a Hudi data source stored in Amazon S3.
- `S3HudiCatalogTarget` – A S3HudiCatalogTarget object.

  Specifies a target that writes to a Hudi data source in the AWS Glue Data Catalog.
- `S3HudiDirectTarget` – A S3HudiDirectTarget object.

  Specifies a target that writes to a Hudi data source in Amazon S3.
- `S3CatalogDeltaSource` – A S3CatalogDeltaSource object.

  Specifies a Delta Lake data source that is registered in the AWS Glue Data Catalog. The data source must be stored in Amazon S3.
- `CatalogDeltaSource` – A CatalogDeltaSource object.

  Specifies a Delta Lake data source that is registered in the AWS Glue Data Catalog.
- `S3DeltaSource` – A S3DeltaSource object.

  Specifies a Delta Lake data source stored in Amazon S3.

- `S3DeltaCatalogTarget` – A S3DeltaCatalogTarget object.

  Specifies a target that writes to a Delta Lake data source in the AWS Glue Data Catalog.
- `S3DeltaDirectTarget` – A S3DeltaDirectTarget object.

  Specifies a target that writes to a Delta Lake data source in Amazon S3.
- `AmazonRedshiftSource` – An AmazonRedshiftSource object.

  Specifies a target that writes to a data source in Amazon Redshift.
- `AmazonRedshiftTarget` – An AmazonRedshiftTarget object.

  Specifies a target that writes to a data target in Amazon Redshift.
- `EvaluateDataQualityMultiFrame` – An EvaluateDataQualityMultiFrame object.

  Specifies your data quality evaluation criteria. Allows multiple input data and returns a collection of Dynamic Frames.
- `Recipe` – A Recipe object.

  Specifies a AWS Glue DataBrew recipe node.
- `SnowflakeSource` – A SnowflakeSource object.

  Specifies a Snowflake data source.
- `SnowflakeTarget` – A SnowflakeTarget object.

  Specifies a target that writes to a Snowflake data source.
- `ConnectorDataSource` – A ConnectorDataSource object.

  Specifies a source generated with standard connection options.
- `ConnectorDataTarget` – A ConnectorDataTarget object.

  Specifies a target generated with standard connection options.
- `S3CatalogIcebergSource` – A S3CatalogIcebergSource object.

  Specifies an Apache Iceberg data source that is registered in the AWS Glue Data Catalog. The Iceberg data source must be stored in Amazon S3.
- `CatalogIcebergSource` – A CatalogIcebergSource object.

  Specifies an Apache Iceberg data source that is registered in the AWS Glue Data Catalog.

- `S3IcebergCatalogTarget` – A S3IcebergCatalogTarget object.

  Specifies an Apache Iceberg catalog target that writes data to Amazon S3 and registers the table in the AWS Glue Data Catalog.
- `S3IcebergDirectTarget` – A S3IcebergDirectTarget object.

  Defines configuration parameters for writing data to Amazon S3 as an Apache Iceberg table.
- `S3ExcelSource` – A S3ExcelSource object.

  Defines configuration parameters for reading Excel files from Amazon S3.
- `S3HyperDirectTarget` – A S3HyperDirectTarget object.

  Defines configuration parameters for writing data to Amazon S3 using HyperDirect optimization.
- `DynamoDBELTConnectorSource` – A DynamoDBELTConnectorSource object.

  Specifies a DynamoDB ELT connector source for extracting data from DynamoDB tables.

JDBCConnectorOptions structure

Additional connection options for the connector.

Fields

- `FilterPredicate` – UTF-8 string, matching the Custom string pattern #59.

  Extra condition clause to filter data from source. For example:

  `BillingCity='Mountain View'`

  When using a query instead of a table name, you should validate that the query works with the specified `filterPredicate`.
- `PartitionColumn` – UTF-8 string, matching the Custom string pattern #59.

  The name of an integer column that is used for partitioning. This option works only when it's included with `lowerBound`, `upperBound`, and `numPartitions`. This option works the same way as in the Spark SQL JDBC reader.
- `LowerBound` – Number (long), not more than None.

  The minimum value of `partitionColumn` that is used to decide partition stride.

- `UpperBound` – Number (long), not more than None.

  The maximum value of `partitionColumn` that is used to decide partition stride.

- `NumPartitions` – Number (long), not more than None.

  The number of partitions. This value, along with `lowerBound` (inclusive) and `upperBound` (exclusive), form partition strides for generated `WHERE` clause expressions that are used to split the `partitionColumn`.

- `JobBookmarkKeys` – An array of UTF-8 strings.

  The name of the job bookmark keys on which to sort.

- `JobBookmarkKeysSortOrder` – UTF-8 string, matching the [Custom string pattern #59](#).

  Specifies an ascending or descending sort order.

- `DataTypeMapping` – A map array of key-value pairs.

  Each key is a UTF-8 string (valid values: `ARRAY` | `BIGINT` | `BINARY` | `BIT` | `BLOB` | `BOOLEAN` | `CHAR` | `CLOB` | `DATALINK` | `DATE` | `DECIMAL` | `DISTINCT` | `DOUBLE` | `FLOAT` | `INTEGER` | `JAVA_OBJECT` | `LONGNVARCHAR` | `LONGVARBINARY` | `LONGVARCHAR` | `NCHAR` | `NCLOB` | `NULL` | `NUMERIC` | `NVARCHAR` | `OTHER` | `REAL` | `REF` | `REF_CURSOR` | `ROWID` | `SMALLINT` | `SQLXML` | `STRUCT` | `TIME` | `TIME_WITH_TIMEZONE` | `TIMESTAMP` | `TIMESTAMP_WITH_TIMEZONE` | `TINYINT` | `VARBINARY` | `VARCHAR`).

  Each value is a UTF-8 string (valid values: `DATE` | `STRING` | `TIMESTAMP` | `INT` | `FLOAT` | `LONG` | `BIGDECIMAL` | `BYTE` | `SHORT` | `DOUBLE`).

  Custom data type mapping that builds a mapping from a JDBC data type to an AWS Glue data type. For example, the option `"dataTypeMapping":{"FLOAT":"STRING"}` maps data fields of JDBC type `FLOAT` into the Java `String` type by calling the `ResultSet.getString()` method of the driver, and uses it to build the AWS Glue record. The `ResultSet` object is implemented by each driver, so the behavior is specific to the driver you use. Refer to the documentation for your JDBC driver to understand how the driver performs the conversions.

StreamingDataPreviewOptions structure

Specifies options related to data preview for viewing a sample of your data.

Fields

- `PollingTime` – Number (long), at least 10.

  The polling time in milliseconds.
- `RecordPollingLimit` – Number (long), at least 1.

  The limit to the number of records polled.

AthenaConnectorSource structure

Specifies a connector to an Amazon Athena data source.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data source.
- `ConnectionName` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the connection that is associated with the connector.
- `ConnectorName` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of a connector that assists with accessing the data store in AWS Glue Studio.
- `ConnectionType` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The type of connection, such as marketplace.athena or custom.athena, designating a connection to an Amazon Athena data store.
- `ConnectionTable` – UTF-8 string, matching the Custom string pattern #59.

  The name of the table in the data source.
- `SchemaName` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the Cloudwatch log group to read from. For example, `/aws-glue/jobs/output`.
- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for the custom Athena source.

JDBCConnectorSource structure

Specifies a connector to a JDBC data source.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data source.
- `ConnectionName` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the connection that is associated with the connector.
- `ConnectorName` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of a connector that assists with accessing the data store in AWS Glue Studio.
- `ConnectionType` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The type of connection, such as marketplace.jdbc or custom.jdbc, designating a connection to a JDBC data store.
- `AdditionalOptions` – A JDBCConnectorOptions object.

  Additional connection options for the connector.
- `ConnectionTable` – UTF-8 string, matching the Custom string pattern #59.

  The name of the table in the data source.
- `Query` – UTF-8 string, matching the Custom string pattern #60.

  The table or SQL query to get the data from. You can specify either `ConnectionTable` or `query`, but not both.
- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for the custom JDBC source.

SparkConnectorSource structure

Specifies a connector to an Apache Spark data source.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data source.
- `ConnectionName` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the connection that is associated with the connector.
- `ConnectorName` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of a connector that assists with accessing the data store in AWS Glue Studio.
- `ConnectionType` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The type of connection, such as marketplace.spark or custom.spark, designating a connection to an Apache Spark data store.
- `AdditionalOptions` – A map array of key-value pairs.

  Each key is a UTF-8 string, matching the Custom string pattern #59.

  Each value is a UTF-8 string, matching the Custom string pattern #59.

  Additional connection options for the connector.
- `OutputSchemas` – An array of GlueSchema objects.

  Specifies data schema for the custom spark source.

CatalogSource structure

Specifies a data store in the AWS Glue Data Catalog.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data store.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the database to read from.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

The name of the table in the database to read from.

- `PartitionPredicate` – UTF-8 string, matching the Custom string pattern #59.

  Partitions satisfying this predicate are deleted. Files within the retention period in these partitions are not deleted.
- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for the catalog source.

MySQLCatalogSource structure

Specifies a MySQL data source in the AWS Glue Data Catalog.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data source.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the database to read from.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the table in the database to read from.

PostgreSQLCatalogSource structure

Specifies a PostgresSQL data source in the AWS Glue Data Catalog.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data source.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the database to read from.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

The name of the table in the database to read from.

OracleSQLCatalogSource structure

Specifies an Oracle data source in the AWS Glue Data Catalog.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data source.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the database to read from.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the table in the database to read from.

MicrosoftSQLServerCatalogSource structure

Specifies a Microsoft SQL server data source in the AWS Glue Data Catalog.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data source.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the database to read from.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the table in the database to read from.

CatalogKinesisSource structure

Specifies a Kinesis data source in the AWS Glue Data Catalog.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data source.
- `WindowSize` – Number (integer), not more than None.

  The amount of time to spend processing each micro batch.
- `DetectSchema` – Boolean.

  Whether to automatically determine the schema from the incoming data.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the table in the database to read from.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the database to read from.
- `StreamingOptions` – A KinesisStreamingSourceOptions object.

  Additional options for the Kinesis streaming data source.
- `DataPreviewOptions` – A StreamingDataPreviewOptions object.

  Additional options for data preview.

DirectKinesisSource structure

Specifies a direct Amazon Kinesis data source.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data source.
- `WindowSize` – Number (integer), not more than None.

  The amount of time to spend processing each micro batch.
- `DetectSchema` – Boolean.

  Whether to automatically determine the schema from the incoming data.

- `StreamingOptions` – A KinesisStreamingSourceOptions object.

  Additional options for the Kinesis streaming data source.
- `DataPreviewOptions` – A StreamingDataPreviewOptions object.

  Additional options for data preview.

KinesisStreamingSourceOptions structure

Additional options for the Amazon Kinesis streaming data source.

Fields

- `EndpointUrl` – UTF-8 string, matching the Custom string pattern #59.

  The URL of the Kinesis endpoint.
- `StreamName` – UTF-8 string, matching the Custom string pattern #59.

  The name of the Kinesis data stream.
- `Classification` – UTF-8 string, matching the Custom string pattern #59.

  An optional classification.
- `Delimiter` – UTF-8 string, matching the Custom string pattern #59.

  Specifies the delimiter character.
- `StartingPosition` – UTF-8 string (valid values: `latest="LATEST"` | `trim_horizon="TRIM_HORIZON"` | `earliest="EARLIEST"` | `timestamp="TIMESTAMP"`).

  The starting position in the Kinesis data stream to read data from. The possible values are `"latest"`, `"trim_horizon"`, `"earliest"`, or a timestamp string in UTC format in the pattern `yyyy-mm-ddTHH:MM:SSZ` (where `Z` represents a UTC timezone offset with a +/-. For example: "2023-04-04T08:00:00-04:00"). The default value is `"latest"`.

  Note: Using a value that is a timestamp string in UTC format for "startingPosition" is supported only for AWS Glue version 4.0 or later.
- `MaxFetchTimeInMs` – Number (long), not more than None.

The maximum time spent for the job executor to read records for the current batch from the Kinesis data stream, specified in milliseconds (ms). Multiple `GetRecords` API calls may be made within this time. The default value is `1000`.

- `MaxFetchRecordsPerShard` – Number (long), not more than None.

  The maximum number of records to fetch per shard in the Kinesis data stream per microbatch. Note: The client can exceed this limit if the streaming job has already read extra records from Kinesis (in the same get-records call). If `MaxFetchRecordsPerShard` needs to be strict then it needs to be a multiple of `MaxRecordPerRead`. The default value is `100000`.
- `MaxRecordPerRead` – Number (long), not more than None.

  The maximum number of records to fetch from the Kinesis data stream in each getRecords operation. The default value is `10000`.
- `AddIdleTimeBetweenReads` – Boolean.

  Adds a time delay between two consecutive getRecords operations. The default value is `"False"`. This option is only configurable for AWS Glue version 2.0 and above.
- `IdleTimeBetweenReadsInMs` – Number (long), not more than None.

  The minimum time delay between two consecutive getRecords operations, specified in ms. The default value is `1000`. This option is only configurable for AWS Glue version 2.0 and above.
- `DescribeShardInterval` – Number (long), not more than None.

  The minimum time interval between two ListShards API calls for your script to consider resharding. The default value is `1s`.
- `NumRetries` – Number (integer), not more than None.

  The maximum number of retries for Kinesis Data Streams API requests. The default value is `3`.
- `RetryIntervalMs` – Number (long), not more than None.

  The cool-off time period (specified in ms) before retrying the Kinesis Data Streams API call. The default value is `1000`.
- `MaxRetryIntervalMs` – Number (long), not more than None.

  The maximum cool-off time period (specified in ms) between two retries of a Kinesis Data Streams API call. The default value is `10000`.

- `AvoidEmptyBatches` – Boolean.

  Avoids creating an empty microbatch job by checking for unread data in the Kinesis data stream before the batch is started. The default value is `"False"`.
- `StreamArn` – UTF-8 string, matching the Custom string pattern #59.

  The Amazon Resource Name (ARN) of the Kinesis data stream.
- `RoleArn` – UTF-8 string, matching the Custom string pattern #59.

  The Amazon Resource Name (ARN) of the role to assume using AWS Security Token Service (AWS STS). This role must have permissions for describe or read record operations for the Kinesis data stream. You must use this parameter when accessing a data stream in a different account. Used in conjunction with `"awsSTSSessionName"`.
- `RoleSessionName` – UTF-8 string, matching the Custom string pattern #59.

  An identifier for the session assuming the role using AWS STS. You must use this parameter when accessing a data stream in a different account. Used in conjunction with `"awsSTSRoleARN"`.
- `AddRecordTimestamp` – UTF-8 string, matching the Custom string pattern #59.

  When this option is set to 'true', the data output will contain an additional column named "__src_timestamp" that indicates the time when the corresponding record received by the stream. The default value is 'false'. This option is supported in AWS Glue version 4.0 or later.
- `EmitConsumerLagMetrics` – UTF-8 string, matching the Custom string pattern #59.

  When this option is set to 'true', for each batch, it will emit the metrics for the duration between the oldest record received by the stream and the time it arrives in AWS Glue to CloudWatch. The metric's name is "glue.driver.streaming.maxConsumerLagInMs". The default value is 'false'. This option is supported in AWS Glue version 4.0 or later.
- `StartingTimestamp` – UTF-8 string.

  The timestamp of the record in the Kinesis data stream to start reading data from. The possible values are a timestamp string in UTC format of the pattern `yyyy-mm-ddTHH:MM:SSZ` (where Z represents a UTC timezone offset with a +/-. For example: "2023-04-04T08:00:00+08:00").
- `FanoutConsumerARN` – UTF-8 string, matching the Custom string pattern #59.

The Amazon Resource Name (ARN) of the Kinesis Data Streams enhanced fan-out consumer. When specified, enables enhanced fan-out for dedicated throughput and lower latency data consumption.

CatalogKafkaSource structure

Specifies an Apache Kafka data store in the Data Catalog.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data store.
- `WindowSize` – Number (integer), not more than None.

  The amount of time to spend processing each micro batch.
- `DetectSchema` – Boolean.

  Whether to automatically determine the schema from the incoming data.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the table in the database to read from.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the database to read from.
- `StreamingOptions` – A KafkaStreamingSourceOptions object.

  Specifies the streaming options.
- `DataPreviewOptions` – A StreamingDataPreviewOptions object.

  Specifies options related to data preview for viewing a sample of your data.

DirectKafkaSource structure

Specifies an Apache Kafka data store.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data store.
- `StreamingOptions` – A KafkaStreamingSourceOptions object.

  Specifies the streaming options.
- `WindowSize` – Number (integer), not more than None.

  The amount of time to spend processing each micro batch.
- `DetectSchema` – Boolean.

  Whether to automatically determine the schema from the incoming data.
- `DataPreviewOptions` – A StreamingDataPreviewOptions object.

  Specifies options related to data preview for viewing a sample of your data.

KafkaStreamingSourceOptions structure

Additional options for streaming.

Fields

- `BootstrapServers` – UTF-8 string, matching the Custom string pattern #59.

  A list of bootstrap server URLs, for example, as `b-1.vpc-test-2.o4q88o.c6.kafka.us-east-1.amazonaws.com:9094`. This option must be specified in the API call or defined in the table metadata in the Data Catalog.
- `SecurityProtocol` – UTF-8 string, matching the Custom string pattern #59.

  The protocol used to communicate with brokers. The possible values are `"SSL"` or `"PLAINTEXT"`.
- `ConnectionName` – UTF-8 string, matching the Custom string pattern #59.

  The name of the connection.
- `TopicName` – UTF-8 string, matching the Custom string pattern #59.

The topic name as specified in Apache Kafka. You must specify at least one of `"topicName"`, `"assign"` or `"subscribePattern"`.

- `Assign` – UTF-8 string, matching the Custom string pattern #59.

  The specific `TopicPartitions` to consume. You must specify at least one of `"topicName"`, `"assign"` or `"subscribePattern"`.

- `SubscribePattern` – UTF-8 string, matching the Custom string pattern #59.

  A Java regex string that identifies the topic list to subscribe to. You must specify at least one of `"topicName"`, `"assign"` or `"subscribePattern"`.

- `Classification` – UTF-8 string, matching the Custom string pattern #59.

  An optional classification.

- `Delimiter` – UTF-8 string, matching the Custom string pattern #59.

  Specifies the delimiter character.

- `StartingOffsets` – UTF-8 string, matching the Custom string pattern #59.

  The starting position in the Kafka topic to read data from. The possible values are `"earliest"` or `"latest"`. The default value is `"latest"`.

- `EndingOffsets` – UTF-8 string, matching the Custom string pattern #59.

  The end point when a batch query is ended. Possible values are either `"latest"` or a JSON string that specifies an ending offset for each `TopicPartition`.

- `PollTimeoutMs` – Number (long), not more than None.

  The timeout in milliseconds to poll data from Kafka in Spark job executors. The default value is `512`.

- `NumRetries` – Number (integer), not more than None.

  The number of times to retry before failing to fetch Kafka offsets. The default value is `3`.

- `RetryIntervalMs` – Number (long), not more than None.

  The time in milliseconds to wait before retrying to fetch Kafka offsets. The default value is `10`.

- `MaxOffsetsPerTrigger` – Number (long), not more than None.

The rate limit on the maximum number of offsets that are processed per trigger interval. The specified total number of offsets is proportionally split across `topicPartitions` of different volumes. The default value is null, which means that the consumer reads all offsets until the known latest offset.

- `MinPartitions` – Number (integer), not more than None.

  The desired minimum number of partitions to read from Kafka. The default value is null, which means that the number of spark partitions is equal to the number of Kafka partitions.
- `IncludeHeaders` – Boolean.

  Whether to include the Kafka headers. When the option is set to "true", the data output will contain an additional column named "glue_streaming_kafka_headers" with type `Array[Struct(key: String, value: String)]`. The default value is "false". This option is available in AWS Glue version 3.0 or later only.
- `AddRecordTimestamp` – UTF-8 string, matching the Custom string pattern #59.

  When this option is set to 'true', the data output will contain an additional column named "__src_timestamp" that indicates the time when the corresponding record received by the topic. The default value is 'false'. This option is supported in AWS Glue version 4.0 or later.
- `EmitConsumerLagMetrics` – UTF-8 string, matching the Custom string pattern #59.

  When this option is set to 'true', for each batch, it will emit the metrics for the duration between the oldest record received by the topic and the time it arrives in AWS Glue to CloudWatch. The metric's name is "glue.driver.streaming.maxConsumerLagInMs". The default value is 'false'. This option is supported in AWS Glue version 4.0 or later.
- `StartingTimestamp` – UTF-8 string.

  The timestamp of the record in the Kafka topic to start reading data from. The possible values are a timestamp string in UTC format of the pattern `yyyy-mm-ddTHH:MM:SSZ` (where Z represents a UTC timezone offset with a +/-. For example: "2023-04-04T08:00:00+08:00").

  Only one of `StartingTimestamp` or `StartingOffsets` must be set.

RedshiftSource structure

Specifies an Amazon Redshift data store.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the Amazon Redshift data store.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The database to read from.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The database table to read from.
- `RedshiftTmpDir` – UTF-8 string, matching the Custom string pattern #59.

  The Amazon S3 path where temporary data can be staged when copying out of the database.
- `TmpDirIAMRole` – UTF-8 string, matching the Custom string pattern #59.

  The IAM role with permissions.

AmazonRedshiftSource structure

Specifies an Amazon Redshift source.

Fields

- `Name` – UTF-8 string, matching the Custom string pattern #61.

  The name of the Amazon Redshift source.
- `Data` – An AmazonRedshiftNodeData object.

  Specifies the data of the Amazon Reshift source node.

AmazonRedshiftNodeData structure

Specifies an Amazon Redshift node.

Fields

- `AccessType` – UTF-8 string, matching the Custom string pattern #58.

  The access type for the Redshift connection. Can be a direct connection or catalog connections.

- `SourceType` – UTF-8 string, matching the Custom string pattern #58.

  The source type to specify whether a specific table is the source or a custom query.
- `Connection` – An Option object.

  The AWS Glue connection to the Redshift cluster.
- `Schema` – An Option object.

  The Redshift schema name when working with a direct connection.
- `Table` – An Option object.

  The Redshift table name when working with a direct connection.
- `CatalogDatabase` – An Option object.

  The name of the AWS Glue Data Catalog database when working with a data catalog.
- `CatalogTable` – An Option object.

  The AWS Glue Data Catalog table name when working with a data catalog.
- `CatalogRedshiftSchema` – UTF-8 string.

  The Redshift schema name when working with a data catalog.
- `CatalogRedshiftTable` – UTF-8 string.

  The database table to read from.
- `TempDir` – UTF-8 string, matching the Custom string pattern #59.

  The Amazon S3 path where temporary data can be staged when copying out of the database.
- `IamRole` – An Option object.

  Optional. The role name use when connection to S3. The IAM role ill default to the role on the job when left blank.
- `AdvancedOptions` – An array of AmazonRedshiftAdvancedOption objects.

  Optional values when connecting to the Redshift cluster.
- `SampleQuery` – UTF-8 string.

  The SQL used to fetch the data from a Redshift sources when the SourceType is 'query'.
- `PreAction` – UTF-8 string.

The SQL used before a MERGE or APPEND with upsert is run.

- `PostAction` – UTF-8 string.

  The SQL used before a MERGE or APPEND with upsert is run.

- `Action` – UTF-8 string.

  Specifies how writing to a Redshift cluser will occur.

- `TablePrefix` – UTF-8 string, matching the Custom string pattern #58.

  Specifies the prefix to a table.

- `Upsert` – Boolean.

  The action used on Redshift sinks when doing an APPEND.

- `MergeAction` – UTF-8 string, matching the Custom string pattern #58.

  The action used when to detemine how a MERGE in a Redshift sink will be handled.

- `MergeWhenMatched` – UTF-8 string, matching the Custom string pattern #58.

  The action used when to detemine how a MERGE in a Redshift sink will be handled when an existing record matches a new record.

- `MergeWhenNotMatched` – UTF-8 string, matching the Custom string pattern #58.

  The action used when to detemine how a MERGE in a Redshift sink will be handled when an existing record doesn't match a new record.

- `MergeClause` – UTF-8 string.

  The SQL used in a custom merge to deal with matching records.

- `CrawlerConnection` – UTF-8 string.

  Specifies the name of the connection that is associated with the catalog table used.

- `TableSchema` – An array of Option objects.

  The array of schema output for a given node.

- `StagingTable` – UTF-8 string.

  The name of the temporary staging table that is used when doing a MERGE or APPEND with upsert.

- `SelectedColumns` – An array of Option objects.

  The list of column names used to determine a matching record when doing a MERGE or APPEND with upsert.

AmazonRedshiftAdvancedOption structure

Specifies an optional value when connecting to the Redshift cluster.

Fields

- `Key` – UTF-8 string.

  The key for the additional connection option.
- `Value` – UTF-8 string.

  The value for the additional connection option.

Option structure

Specifies an option value.

Fields

- `Value` – UTF-8 string, matching the Custom string pattern #59.

  Specifies the value of the option.
- `Label` – UTF-8 string, matching the Custom string pattern #59.

  Specifies the label of the option.
- `Description` – UTF-8 string, matching the Custom string pattern #59.

  Specifies the description of the option.

S3CatalogSource structure

Specifies an Amazon S3 data store in the AWS Glue Data Catalog.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data store.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The database to read from.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The database table to read from.
- `PartitionPredicate` – UTF-8 string, matching the Custom string pattern #59.

  Partitions satisfying this predicate are deleted. Files within the retention period in these partitions are not deleted. Set to `""` – empty by default.
- `AdditionalOptions` – A S3SourceAdditionalOptions object.

  Specifies additional connection options.

S3SourceAdditionalOptions structure

Specifies additional connection options for the Amazon S3 data store.

Fields

- `BoundedSize` – Number (long).

  Sets the upper limit for the target size of the dataset in bytes that will be processed.
- `BoundedFiles` – Number (long).

  Sets the upper limit for the target number of files that will be processed.

S3CsvSource structure

Specifies a command-separated value (CSV) data store stored in Amazon S3.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

The name of the data store.

- `Paths` – *Required:* An array of UTF-8 strings.

  A list of the Amazon S3 paths to read from.
- `CompressionType` – UTF-8 string (valid values: `gzip="GZIP"` | `bzip2="BZIP2"`).

  Specifies how the data is compressed. This is generally not necessary if the data has a standard file extension. Possible values are `"gzip"` and `"bzip"`).
- `Exclusions` – An array of UTF-8 strings.

  A string containing a JSON list of Unix-style glob patterns to exclude. For example, "[\"**.pdf\"]" excludes all PDF files.
- `GroupSize` – UTF-8 string, matching the Custom string pattern #59.

  The target group size in bytes. The default is computed based on the input data size and the size of your cluster. When there are fewer than 50,000 input files, `"groupFiles"` must be set to `"inPartition"` for this to take effect.
- `GroupFiles` – UTF-8 string, matching the Custom string pattern #59.

  Grouping files is turned on by default when the input contains more than 50,000 files. To turn on grouping with fewer than 50,000 files, set this parameter to "inPartition". To disable grouping when there are more than 50,000 files, set this parameter to `"none"`.
- `Recurse` – Boolean.

  If set to true, recursively reads files in all subdirectories under the specified paths.
- `MaxBand` – Number (integer), not more than None.

  This option controls the duration in milliseconds after which the s3 listing is likely to be consistent. Files with modification timestamps falling within the last maxBand milliseconds are tracked specially when using JobBookmarks to account for Amazon S3 eventual consistency. Most users don't need to set this option. The default is 900000 milliseconds, or 15 minutes.
- `MaxFilesInBand` – Number (integer), not more than None.

  This option specifies the maximum number of files to save from the last maxBand seconds. If this number is exceeded, extra files are skipped and only processed in the next job run.
- `AdditionalOptions` – A S3DirectSourceAdditionalOptions object.

Specifies additional connection options.

- `Separator` – *Required:* UTF-8 string (valid values: `comma="COMMA"` | `ctrla="CTRLA"` | `pipe="PIPE"` | `semicolon="SEMICOLON"` | `tab="TAB"`).

  Specifies the delimiter character. The default is a comma: ",", but any other character can be specified.
- `Escaper` – UTF-8 string, matching the Custom string pattern #59.

  Specifies a character to use for escaping. This option is used only when reading CSV files. The default value is `none`. If enabled, the character which immediately follows is used as-is, except for a small set of well-known escapes (`\n`, `\r`, `\t`, and `\0`).
- `QuoteChar` – *Required:* UTF-8 string (valid values: `quote="QUOTE"` | `quillemet="QUILLEMET"` | `single_quote="SINGLE_QUOTE"` | `disabled="DISABLED"`).

  Specifies the character to use for quoting. The default is a double quote: `'"'`. Set this to `-1` to turn off quoting entirely.
- `Multiline` – Boolean.

  A Boolean value that specifies whether a single record can span multiple lines. This can occur when a field contains a quoted new-line character. You must set this option to True if any record spans multiple lines. The default value is `False`, which allows for more aggressive file-splitting during parsing.
- `WithHeader` – Boolean.

  A Boolean value that specifies whether to treat the first line as a header. The default value is `False`.
- `WriteHeader` – Boolean.

  A Boolean value that specifies whether to write the header to output. The default value is `True`.
- `SkipFirst` – Boolean.

  A Boolean value that specifies whether to skip the first data line. The default value is `False`.
- `OptimizePerformance` – Boolean.

  A Boolean value that specifies whether to use the advanced SIMD CSV reader along with Apache Arrow based columnar memory formats. Only available in AWS Glue version 3.0.

- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for the S3 CSV source.

DirectJDBCSource structure

Specifies the direct JDBC source connection.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the JDBC source connection.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The database of the JDBC source connection.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The table of the JDBC source connection.
- `ConnectionName` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The connection name of the JDBC source.
- `ConnectionType` – *Required:* UTF-8 string (valid values: `sqlserver` | `mysql` | `oracle` | `postgresql` | `redshift`).

  The connection type of the JDBC source.
- `RedshiftTmpDir` – UTF-8 string, matching the Custom string pattern #59.

  The temp directory of the JDBC Redshift source.
- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for the direct JDBC source.

S3DirectSourceAdditionalOptions structure

Specifies additional connection options for the Amazon S3 data store.

Fields

- `BoundedSize` – Number (long).

  Sets the upper limit for the target size of the dataset in bytes that will be processed.
- `BoundedFiles` – Number (long).

  Sets the upper limit for the target number of files that will be processed.
- `EnableSamplePath` – Boolean.

  Sets option to enable a sample path.
- `SamplePath` – UTF-8 string, matching the Custom string pattern #59.

  If enabled, specifies the sample path.

S3JsonSource structure

Specifies a JSON data store stored in Amazon S3.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data store.
- `Paths` – *Required:* An array of UTF-8 strings.

  A list of the Amazon S3 paths to read from.
- `CompressionType` – UTF-8 string (valid values: `gzip="GZIP"` | `bzip2="BZIP2"`).

  Specifies how the data is compressed. This is generally not necessary if the data has a standard file extension. Possible values are `"gzip"` and `"bzip"`).
- `Exclusions` – An array of UTF-8 strings.

  A string containing a JSON list of Unix-style glob patterns to exclude. For example, "[\"**.pdf\"]" excludes all PDF files.
- `GroupSize` – UTF-8 string, matching the Custom string pattern #59.

The target group size in bytes. The default is computed based on the input data size and the size of your cluster. When there are fewer than 50,000 input files, `"groupFiles"` must be set to `"inPartition"` for this to take effect.

- `GroupFiles` – UTF-8 string, matching the Custom string pattern #59.

  Grouping files is turned on by default when the input contains more than 50,000 files. To turn on grouping with fewer than 50,000 files, set this parameter to "inPartition". To disable grouping when there are more than 50,000 files, set this parameter to `"none"`.

- `Recurse` – Boolean.

  If set to true, recursively reads files in all subdirectories under the specified paths.

- `MaxBand` – Number (integer), not more than None.

  This option controls the duration in milliseconds after which the s3 listing is likely to be consistent. Files with modification timestamps falling within the last maxBand milliseconds are tracked specially when using JobBookmarks to account for Amazon S3 eventual consistency. Most users don't need to set this option. The default is 900000 milliseconds, or 15 minutes.

- `MaxFilesInBand` – Number (integer), not more than None.

  This option specifies the maximum number of files to save from the last maxBand seconds. If this number is exceeded, extra files are skipped and only processed in the next job run.

- `AdditionalOptions` – A S3DirectSourceAdditionalOptions object.

  Specifies additional connection options.

- `JsonPath` – UTF-8 string, matching the Custom string pattern #59.

  A JsonPath string defining the JSON data.

- `Multiline` – Boolean.

  A Boolean value that specifies whether a single record can span multiple lines. This can occur when a field contains a quoted new-line character. You must set this option to True if any record spans multiple lines. The default value is `False`, which allows for more aggressive file-splitting during parsing.

- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for the S3 JSON source.

S3ParquetSource structure

Specifies an Apache Parquet data store stored in Amazon S3.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data store.

- `Paths` – *Required:* An array of UTF-8 strings.

  A list of the Amazon S3 paths to read from.

- `CompressionType` – UTF-8 string (valid values: `snappy="SNAPPY"` | `lzo="LZO"` | `gzip="GZIP"` | `brotli="BROTLI"` | `lz4="LZ4"` | `uncompressed="UNCOMPRESSED"` | `none="NONE"`).

  Specifies how the data is compressed. This is generally not necessary if the data has a standard file extension. Possible values are `"gzip"` and `"bzip"`).

- `Exclusions` – An array of UTF-8 strings.

  A string containing a JSON list of Unix-style glob patterns to exclude. For example, "[\"**.pdf\"]" excludes all PDF files.

- `GroupSize` – UTF-8 string, matching the Custom string pattern #59.

  The target group size in bytes. The default is computed based on the input data size and the size of your cluster. When there are fewer than 50,000 input files, `"groupFiles"` must be set to `"inPartition"` for this to take effect.

- `GroupFiles` – UTF-8 string, matching the Custom string pattern #59.

  Grouping files is turned on by default when the input contains more than 50,000 files. To turn on grouping with fewer than 50,000 files, set this parameter to "inPartition". To disable grouping when there are more than 50,000 files, set this parameter to `"none"`.

- `Recurse` – Boolean.

  If set to true, recursively reads files in all subdirectories under the specified paths.

- `MaxBand` – Number (integer), not more than None.

This option controls the duration in milliseconds after which the s3 listing is likely to be consistent. Files with modification timestamps falling within the last maxBand milliseconds are tracked specially when using JobBookmarks to account for Amazon S3 eventual consistency. Most users don't need to set this option. The default is 900000 milliseconds, or 15 minutes.

- `MaxFilesInBand` – Number (integer), not more than None.

  This option specifies the maximum number of files to save from the last maxBand seconds. If this number is exceeded, extra files are skipped and only processed in the next job run.
- `AdditionalOptions` – A S3DirectSourceAdditionalOptions object.

  Specifies additional connection options.
- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for the S3 Parquet source.

S3DeltaSource structure

Specifies a Delta Lake data source stored in Amazon S3.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the Delta Lake source.
- `Paths` – *Required:* An array of UTF-8 strings.

  A list of the Amazon S3 paths to read from.
- `AdditionalDeltaOptions` – A map array of key-value pairs.

  Each key is a UTF-8 string, matching the Custom string pattern #59.

  Each value is a UTF-8 string, matching the Custom string pattern #59.

  Specifies additional connection options.
- `AdditionalOptions` – A S3DirectSourceAdditionalOptions object.

  Specifies additional options for the connector.
- `OutputSchemas` – An array of GlueSchema objects.

Specifies the data schema for the Delta Lake source.

S3CatalogDeltaSource structure

Specifies a Delta Lake data source that is registered in the AWS Glue Data Catalog. The data source must be stored in Amazon S3.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the Delta Lake data source.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the database to read from.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the table in the database to read from.
- `AdditionalDeltaOptions` – A map array of key-value pairs.

  Each key is a UTF-8 string, matching the Custom string pattern #59.

  Each value is a UTF-8 string, matching the Custom string pattern #59.

  Specifies additional connection options.
- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for the Delta Lake source.

CatalogDeltaSource structure

Specifies a Delta Lake data source that is registered in the AWS Glue Data Catalog.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the Delta Lake data source.

- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the database to read from.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the table in the database to read from.
- `AdditionalDeltaOptions` – A map array of key-value pairs.

  Each key is a UTF-8 string, matching the Custom string pattern #59.

  Each value is a UTF-8 string, matching the Custom string pattern #59.

  Specifies additional connection options.
- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for the Delta Lake source.

S3HudiSource structure

Specifies a Hudi data source stored in Amazon S3.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the Hudi source.
- `Paths` – *Required:* An array of UTF-8 strings.

  A list of the Amazon S3 paths to read from.
- `AdditionalHudiOptions` – A map array of key-value pairs.

  Each key is a UTF-8 string, matching the Custom string pattern #59.

  Each value is a UTF-8 string, matching the Custom string pattern #59.

  Specifies additional connection options.
- `AdditionalOptions` – A S3DirectSourceAdditionalOptions object.

  Specifies additional options for the connector.
- `OutputSchemas` – An array of GlueSchema objects.

Specifies the data schema for the Hudi source.

S3CatalogHudiSource structure

Specifies a Hudi data source that is registered in the AWS Glue Data Catalog. The Hudi data source must be stored in Amazon S3.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the Hudi data source.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the database to read from.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the table in the database to read from.
- `AdditionalHudiOptions` – A map array of key-value pairs.

  Each key is a UTF-8 string, matching the Custom string pattern #59.

  Each value is a UTF-8 string, matching the Custom string pattern #59.

  Specifies additional connection options.
- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for the Hudi source.

S3ExcelSource structure

Specifies an S3 Excel data source.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the S3 Excel data source.

- `Paths` – *Required:* An array of UTF-8 strings.

  The S3 paths where the Excel files are located.
- `CompressionType` – UTF-8 string (valid values: `snappy="SNAPPY"` | `lzo="LZO"` | `gzip="GZIP"` | `brotli="BROTLI"` | `lz4="LZ4"` | `uncompressed="UNCOMPRESSED"` | `none="NONE"`).

  The compression format used for the Excel files.
- `Exclusions` – An array of UTF-8 strings.

  Patterns to exclude specific files or paths from processing.
- `GroupSize` – UTF-8 string, matching the Custom string pattern #59.

  Defines the size of file groups for batch processing.
- `GroupFiles` – UTF-8 string, matching the Custom string pattern #59.

  Specifies how files should be grouped for processing.
- `Recurse` – Boolean.

  Indicates whether to recursively process subdirectories.
- `MaxBand` – Number (integer), not more than None.

  The maximum number of processing bands to use.
- `MaxFilesInBand` – Number (integer), not more than None.

  The maximum number of files to process in each band.
- `AdditionalOptions` – A S3DirectSourceAdditionalOptions object.

  Additional configuration options for S3 direct source processing.
- `NumberRows` – Number (long).

  The number of rows to process from each Excel file.
- `SkipFooter` – Number (integer), not more than None.

  The number of rows to skip at the end of each Excel file.
- `OutputSchemas` – An array of GlueSchema objects.

  The AWS Glue schemas to apply to the processed data.

CatalogHudiSource structure

Specifies a Hudi data source that is registered in the AWS Glue Data Catalog.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the Hudi data source.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the database to read from.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the table in the database to read from.
- `AdditionalHudiOptions` – A map array of key-value pairs.

  Each key is a UTF-8 string, matching the Custom string pattern #59.

  Each value is a UTF-8 string, matching the Custom string pattern #59.

  Specifies additional connection options.
- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for the Hudi source.

DynamoDBCatalogSource structure

Specifies a DynamoDB data source in the AWS Glue Data Catalog.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data source.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the database to read from.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

The name of the table in the database to read from.

- `PitrEnabled` – Boolean.

  Specifies whether Point-in-Time Recovery (PITR) is enabled for the DynamoDB table. When set to `true`, allows reading from a specific point in time. The default value is `false`.
- `AdditionalOptions` – A DDBELTCatalogAdditionalOptions object.

  Specifies additional connection options for the DynamoDB data source.

RelationalCatalogSource structure

Specifies a Relational database data source in the AWS Glue Data Catalog.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data source.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the database to read from.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the table in the database to read from.

JDBCConnectorTarget structure

Specifies a data target that writes to Amazon S3 in Apache Parquet columnar storage.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data target.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The nodes that are inputs to the data target.
- `ConnectionName` – *Required:* UTF-8 string, matching the Custom string pattern #59.

The name of the connection that is associated with the connector.

- `ConnectionTable` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the table in the data target.
- `ConnectorName` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of a connector that will be used.
- `ConnectionType` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The type of connection, such as marketplace.jdbc or custom.jdbc, designating a connection to a JDBC data target.
- `AdditionalOptions` – A map array of key-value pairs.

  Each key is a UTF-8 string, matching the Custom string pattern #59.

  Each value is a UTF-8 string, matching the Custom string pattern #59.

  Additional connection options for the connector.
- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for the JDBC target.

SparkConnectorTarget structure

Specifies a target that uses an Apache Spark connector.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data target.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The nodes that are inputs to the data target.
- `ConnectionName` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of a connection for an Apache Spark connector.
- `ConnectorName` – *Required:* UTF-8 string, matching the Custom string pattern #59.

The name of an Apache Spark connector.

- `ConnectionType` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The type of connection, such as marketplace.spark or custom.spark, designating a connection to an Apache Spark data store.
- `AdditionalOptions` – A map array of key-value pairs.

  Each key is a UTF-8 string, matching the Custom string pattern #59.

  Each value is a UTF-8 string, matching the Custom string pattern #59.

  Additional connection options for the connector.
- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for the custom spark target.

BasicCatalogTarget structure

Specifies a target that uses a AWS Glue Data Catalog table.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of your data target.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The nodes that are inputs to the data target.
- `PartitionKeys` – An array of UTF-8 strings.

  The partition keys used to distribute data across multiple partitions or shards based on a specific key or set of key.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The database that contains the table you want to use as the target. This database must already exist in the Data Catalog.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

The table that defines the schema of your output data. This table must already exist in the Data Catalog.

MySQLCatalogTarget structure

Specifies a target that uses MySQL.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data target.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The nodes that are inputs to the data target.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the database to write to.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the table in the database to write to.

PostgreSQLCatalogTarget structure

Specifies a target that uses Postgres SQL.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data target.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The nodes that are inputs to the data target.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the database to write to.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

The name of the table in the database to write to.

OracleSQLCatalogTarget structure

Specifies a target that uses Oracle SQL.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data target.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The nodes that are inputs to the data target.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the database to write to.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the table in the database to write to.

MicrosoftSQLServerCatalogTarget structure

Specifies a target that uses Microsoft SQL.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data target.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The nodes that are inputs to the data target.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the database to write to.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

The name of the table in the database to write to.

RedshiftTarget structure

Specifies a target that uses Amazon Redshift.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data target.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The nodes that are inputs to the data target.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the database to write to.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the table in the database to write to.
- `RedshiftTmpDir` – UTF-8 string, matching the Custom string pattern #59.

  The Amazon S3 path where temporary data can be staged when copying out of the database.
- `TmpDirIAMRole` – UTF-8 string, matching the Custom string pattern #59.

  The IAM role with permissions.
- `UpsertRedshiftOptions` – An UpsertRedshiftTargetOptions object.

  The set of options to configure an upsert operation when writing to a Redshift target.

AmazonRedshiftTarget structure

Specifies an Amazon Redshift target.

Fields

- `Name` – UTF-8 string, matching the Custom string pattern #61.

The name of the Amazon Redshift target.

- `Data` – An AmazonRedshiftNodeData object.

  Specifies the data of the Amazon Redshift target node.
- `Inputs` – An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The nodes that are inputs to the data target.

UpsertRedshiftTargetOptions structure

The options to configure an upsert operation when writing to a Redshift target .

Fields

- `TableLocation` – UTF-8 string, matching the Custom string pattern #59.

  The physical location of the Redshift table.
- `ConnectionName` – UTF-8 string, matching the Custom string pattern #59.

  The name of the connection to use to write to Redshift.
- `UpsertKeys` – An array of UTF-8 strings.

  The keys used to determine whether to perform an update or insert.

S3CatalogTarget structure

Specifies a data target that writes to Amazon S3 using the AWS Glue Data Catalog.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data target.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The nodes that are inputs to the data target.
- `PartitionKeys` – An array of UTF-8 strings.

Specifies native partitioning using a sequence of keys.

- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the table in the database to write to.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the database to write to.
- `SchemaChangePolicy` – A CatalogSchemaChangePolicy object.

  A policy that specifies update behavior for the crawler.
- `AutoDataQuality` – An AutoDataQuality object.

  Specifies whether to automatically enable data quality evaluation for the S3 catalog target. When set to `true`, data quality checks are performed automatically during the write operation.

S3GlueParquetTarget structure

Specifies a data target that writes to Amazon S3 in Apache Parquet columnar storage.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data target.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The nodes that are inputs to the data target.
- `PartitionKeys` – An array of UTF-8 strings.

  Specifies native partitioning using a sequence of keys.
- `Path` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  A single Amazon S3 path to write to.
- `Compression` – UTF-8 string (valid values: `snappy="SNAPPY"` | `lzo="LZO"` | `gzip="GZIP"` | `brotli="BROTLI"` | `lz4="LZ4"` | `uncompressed="UNCOMPRESSED"` | `none="NONE"`).

  Specifies how the data is compressed. This is generally not necessary if the data has a standard file extension. Possible values are `"gzip"` and `"bzip"`).

- `NumberTargetPartitions` – UTF-8 string.

  Specifies the number of target partitions for Parquet files when writing to Amazon S3 using AWS Glue.
- `SchemaChangePolicy` – A DirectSchemaChangePolicy object.

  A policy that specifies update behavior for the crawler.
- `AutoDataQuality` – An AutoDataQuality object.

  Specifies whether to automatically enable data quality evaluation for the S3 AWS Glue Parquet target. When set to `true`, data quality checks are performed automatically during the write operation.

CatalogSchemaChangePolicy structure

A policy that specifies update behavior for the crawler.

Fields

- `EnableUpdateCatalog` – Boolean.

  Whether to use the specified update behavior when the crawler finds a changed schema.
- `UpdateBehavior` – UTF-8 string (valid values: `UPDATE_IN_DATABASE` | `LOG`).

  The update behavior when the crawler finds a changed schema.

S3DirectTarget structure

Specifies a data target that writes to Amazon S3.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data target.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The nodes that are inputs to the data target.
- `PartitionKeys` – An array of UTF-8 strings.

Specifies native partitioning using a sequence of keys.

- `Path` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  A single Amazon S3 path to write to.
- `Compression` – UTF-8 string, matching the Custom string pattern #59.

  Specifies how the data is compressed. This is generally not necessary if the data has a standard file extension. Possible values are `"gzip"` and `"bzip"`).
- `NumberTargetPartitions` – UTF-8 string.

  Specifies the number of target partitions when writing data directly to Amazon S3.
- `Format` – *Required:* UTF-8 string (valid values: `json="JSON"` | `csv="CSV"` | `avro="AVRO"` | `orc="ORC"` | `parquet="PARQUET"` | `hudi="HUDI"` | `delta="DELTA"` | `iceberg="ICEBERG"` | `hyper="HYPER"` | `xml="XML"`).

  Specifies the data output format for the target.
- `SchemaChangePolicy` – A DirectSchemaChangePolicy object.

  A policy that specifies update behavior for the crawler.
- `AutoDataQuality` – An AutoDataQuality object.

  Specifies whether to automatically enable data quality evaluation for the S3 direct target. When set to `true`, data quality checks are performed automatically during the write operation.
- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for the S3 direct target.

S3HudiCatalogTarget structure

Specifies a target that writes to a Hudi data source in the AWS Glue Data Catalog.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data target.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

The nodes that are inputs to the data target.

- `PartitionKeys` – An array of UTF-8 strings.

  Specifies native partitioning using a sequence of keys.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the table in the database to write to.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the database to write to.
- `AdditionalOptions` – *Required:* A map array of key-value pairs.

  Each key is a UTF-8 string, matching the Custom string pattern #59.

  Each value is a UTF-8 string, matching the Custom string pattern #59.

  Specifies additional connection options for the connector.
- `SchemaChangePolicy` – A CatalogSchemaChangePolicy object.

  A policy that specifies update behavior for the crawler.
- `AutoDataQuality` – An AutoDataQuality object.

  Specifies whether to automatically enable data quality evaluation for the S3 Hudi catalog target. When set to `true`, data quality checks are performed automatically during the write operation.
- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for the S3 Hudi catalog target.

S3HudiDirectTarget structure

Specifies a target that writes to a Hudi data source in Amazon S3.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data target.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

The nodes that are inputs to the data target.

- `Path` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The Amazon S3 path of your Hudi data source to write to.

- `Compression` – *Required:* UTF-8 string (valid values: `gzip="GZIP"` | `lzo="LZO"` | `uncompressed="UNCOMPRESSED"` | `snappy="SNAPPY"`).

  Specifies how the data is compressed. This is generally not necessary if the data has a standard file extension. Possible values are `"gzip"` and `"bzip"`).

- `NumberTargetPartitions` – UTF-8 string.

  Specifies the number of target partitions for distributing Hudi dataset files across Amazon S3.

- `PartitionKeys` – An array of UTF-8 strings.

  Specifies native partitioning using a sequence of keys.

- `Format` – *Required:* UTF-8 string (valid values: `json="JSON"` | `csv="CSV"` | `avro="AVRO"` | `orc="ORC"` | `parquet="PARQUET"` | `hudi="HUDI"` | `delta="DELTA"` | `iceberg="ICEBERG"` | `hyper="HYPER"` | `xml="XML"`).

  Specifies the data output format for the target.

- `AdditionalOptions` – *Required:* A map array of key-value pairs.

  Each key is a UTF-8 string, matching the Custom string pattern #59.

  Each value is a UTF-8 string, matching the Custom string pattern #59.

  Specifies additional connection options for the connector.

- `SchemaChangePolicy` – A DirectSchemaChangePolicy object.

  A policy that specifies update behavior for the crawler.

- `AutoDataQuality` – An AutoDataQuality object.

  Specifies whether to automatically enable data quality evaluation for the S3 Hudi direct target. When set to `true`, data quality checks are performed automatically during the write operation.

S3DeltaCatalogTarget structure

Specifies a target that writes to a Delta Lake data source in the AWS Glue Data Catalog.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data target.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The nodes that are inputs to the data target.
- `PartitionKeys` – An array of UTF-8 strings.

  Specifies native partitioning using a sequence of keys.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the table in the database to write to.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the database to write to.
- `AdditionalOptions` – A map array of key-value pairs.

  Each key is a UTF-8 string, matching the Custom string pattern #59.

  Each value is a UTF-8 string, matching the Custom string pattern #59.

  Specifies additional connection options for the connector.
- `SchemaChangePolicy` – A CatalogSchemaChangePolicy object.

  A policy that specifies update behavior for the crawler.
- `AutoDataQuality` – An AutoDataQuality object.

  Specifies whether to automatically enable data quality evaluation for the S3 Delta catalog target. When set to `true`, data quality checks are performed automatically during the write operation.
- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for the S3 Delta catalog target.

S3DeltaDirectTarget structure

Specifies a target that writes to a Delta Lake data source in Amazon S3.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data target.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The nodes that are inputs to the data target.
- `PartitionKeys` – An array of UTF-8 strings.

  Specifies native partitioning using a sequence of keys.
- `Path` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The Amazon S3 path of your Delta Lake data source to write to.
- `Compression` – *Required:* UTF-8 string (valid values: `uncompressed="UNCOMPRESSED"` | `snappy="SNAPPY"`).

  Specifies how the data is compressed. This is generally not necessary if the data has a standard file extension. Possible values are `"gzip"` and `"bzip"`).
- `NumberTargetPartitions` – UTF-8 string.

  Specifies the number of target partitions for distributing Delta Lake dataset files across Amazon S3.
- `Format` – *Required:* UTF-8 string (valid values: `json="JSON"` | `csv="CSV"` | `avro="AVRO"` | `orc="ORC"` | `parquet="PARQUET"` | `hudi="HUDI"` | `delta="DELTA"` | `iceberg="ICEBERG"` | `hyper="HYPER"` | `xml="XML"`).

  Specifies the data output format for the target.
- `AdditionalOptions` – A map array of key-value pairs.

  Each key is a UTF-8 string, matching the Custom string pattern #59.

  Each value is a UTF-8 string, matching the Custom string pattern #59.

  Specifies additional connection options for the connector.

- `SchemaChangePolicy` – A DirectSchemaChangePolicy object.

  A policy that specifies update behavior for the crawler.
- `AutoDataQuality` – An AutoDataQuality object.

  Specifies whether to automatically enable data quality evaluation for the S3 Delta direct target. When set to `true`, data quality checks are performed automatically during the write operation.

S3HyperDirectTarget structure

Specifies a HyperDirect data target that writes to Amazon S3.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The unique identifier for the HyperDirect target node.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  Specifies the input source for the HyperDirect target.
- `Format` – UTF-8 string (valid values: `json="JSON"` | `csv="CSV"` | `avro="AVRO"` | `orc="ORC"` | `parquet="PARQUET"` | `hudi="HUDI"` | `delta="DELTA"` | `iceberg="ICEBERG"` | `hyper="HYPER"` | `xml="XML"`).

  Specifies the data output format for the HyperDirect target.
- `PartitionKeys` – An array of UTF-8 strings.

  Defines the partitioning strategy for the output data.
- `Path` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The S3 location where the output data will be written.
- `Compression` – UTF-8 string (valid values: `uncompressed="UNCOMPRESSED"`).

  The compression type to apply to the output data.
- `SchemaChangePolicy` – A DirectSchemaChangePolicy object.

  Defines how schema changes are handled during write operations.
- `AutoDataQuality` – An AutoDataQuality object.

Specifies whether to automatically enable data quality evaluation for the S3 Hyper direct target. When set to `true`, data quality checks are performed automatically during the write operation.

- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for the S3 Hyper direct target.

S3IcebergDirectTarget structure

Specifies a target that writes to an Iceberg data source in Amazon S3.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  Specifies the unique identifier for the Iceberg target node in your data pipeline.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  Defines the single input source that provides data to this Iceberg target.
- `PartitionKeys` – An array of UTF-8 strings.

  Specifies the columns used to partition the Iceberg table data in S3.
- `Path` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  Defines the S3 location where the Iceberg table data will be stored.
- `Format` – *Required:* UTF-8 string (valid values: `json="JSON"` | `csv="CSV"` | `avro="AVRO"` | `orc="ORC"` | `parquet="PARQUET"` | `hudi="HUDI"` | `delta="DELTA"` | `iceberg="ICEBERG"` | `hyper="HYPER"` | `xml="XML"`).

  Specifies the file format used for storing Iceberg table data (e.g., Parquet, ORC).
- `AdditionalOptions` – A map array of key-value pairs.

  Each key is a UTF-8 string, matching the Custom string pattern #59.

  Each value is a UTF-8 string, matching the Custom string pattern #59.

  Provides additional configuration options for customizing the Iceberg table behavior.
- `SchemaChangePolicy` – A DirectSchemaChangePolicy object.

Defines how schema changes are handled when writing data to the Iceberg table.

- `Compression` – *Required:* UTF-8 string (valid values: `gzip="GZIP"` | `lzo="LZO"` | `uncompressed="UNCOMPRESSED"` | `snappy="SNAPPY"`).

  Specifies the compression codec used for Iceberg table files in S3.
- `NumberTargetPartitions` – UTF-8 string.

  Sets the number of target partitions for distributing Iceberg table files across S3.
- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for the S3 Iceberg direct target.

DirectSchemaChangePolicy structure

A policy that specifies update behavior for the crawler.

Fields

- `EnableUpdateCatalog` – Boolean.

  Whether to use the specified update behavior when the crawler finds a changed schema.
- `UpdateBehavior` – UTF-8 string (valid values: `UPDATE_IN_DATABASE` | `LOG`).

  The update behavior when the crawler finds a changed schema.
- `Table` – UTF-8 string, matching the Custom string pattern #59.

  Specifies the table in the database that the schema change policy applies to.
- `Database` – UTF-8 string, matching the Custom string pattern #59.

  Specifies the database that the schema change policy applies to.

ApplyMapping structure

Specifies a transform that maps data property keys in the data source to data property keys in the data target. You can rename keys, modify the data types for keys, and choose which keys to drop from the dataset.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the transform node.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The data inputs identified by their node names.
- `Mapping` – *Required:* An array of Mapping objects.

  Specifies the mapping of data property keys in the data source to data property keys in the data target.

Mapping structure

Specifies the mapping of data property keys.

Fields

- `ToKey` – UTF-8 string, matching the Custom string pattern #59.

  After the apply mapping, what the name of the column should be. Can be the same as `FromPath`.
- `FromPath` – An array of UTF-8 strings.

  The table or column to be modified.
- `FromType` – UTF-8 string, matching the Custom string pattern #59.

  The type of the data to be modified.
- `ToType` – UTF-8 string, matching the Custom string pattern #59.

  The data type that the data is to be modified to.
- `Dropped` – Boolean.

  If true, then the column is removed.
- `Children` – An array of Mapping objects.

Only applicable to nested data structures. If you want to change the parent structure, but also one of its children, you can fill out this data strucutre. It is also `Mapping`, but its `FromPath` will be the parent's `FromPath` plus the `FromPath` from this structure.

For the children part, suppose you have the structure:

```
{ "FromPath": "OuterStructure", "ToKey": "OuterStructure", "ToType":
"Struct", "Dropped": false, "Chidlren": [{ "FromPath": "inner", "ToKey":
"inner", "ToType": "Double", "Dropped": false, }] }
```

You can specify a `Mapping` that looks like:

```
{ "FromPath": "OuterStructure", "ToKey": "OuterStructure", "ToType":
"Struct", "Dropped": false, "Chidlren": [{ "FromPath": "inner", "ToKey":
"inner", "ToType": "Double", "Dropped": false, }] }
```

SelectFields structure

Specifies a transform that chooses the data property keys that you want to keep.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the transform node.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The data inputs identified by their node names.
- `Paths` – *Required:* An array of UTF-8 strings.

  A JSON path to a variable in the data structure.

DropFields structure

Specifies a transform that chooses the data property keys that you want to drop.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

The name of the transform node.

- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The data inputs identified by their node names.
- `Paths` – *Required:* An array of UTF-8 strings.

  A JSON path to a variable in the data structure.

RenameField structure

Specifies a transform that renames a single data property key.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the transform node.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The data inputs identified by their node names.
- `SourcePath` – *Required:* An array of UTF-8 strings.

  A JSON path to a variable in the data structure for the source data.
- `TargetPath` – *Required:* An array of UTF-8 strings.

  A JSON path to a variable in the data structure for the target data.

Spigot structure

Specifies a transform that writes samples of the data to an Amazon S3 bucket.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the transform node.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

The data inputs identified by their node names.

- `Path` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  A path in Amazon S3 where the transform will write a subset of records from the dataset to a JSON file in an Amazon S3 bucket.
- `Topk` – Number (integer), not more than 100.

  Specifies a number of records to write starting from the beginning of the dataset.
- `Prob` – Number (double), not more than 1.

  The probability (a decimal value with a maximum value of 1) of picking any given record. A value of 1 indicates that each row read from the dataset should be included in the sample output.

Join structure

Specifies a transform that joins two datasets into one dataset using a comparison phrase on the specified data property keys. You can use inner, outer, left, right, left semi, and left anti joins.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the transform node.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 2 or more than 2 strings.

  The data inputs identified by their node names.
- `JoinType` – *Required:* UTF-8 string (valid values: `equijoin="EQUIJOIN"` | `left="LEFT"` | `right="RIGHT"` | `outer="OUTER"` | `leftsemi="LEFT_SEMI"` | `leftanti="LEFT_ANTI"`).

  Specifies the type of join to be performed on the datasets.
- `Columns` – *Required:* An array of JoinColumn objects, not less than 2 or more than 2 structures.

  A list of the two columns to be joined.

JoinColumn structure

Specifies a column to be joined.

Fields

- `From` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The column to be joined.
- `Keys` – *Required:* An array of UTF-8 strings.

  The key of the column to be joined.

SplitFields structure

Specifies a transform that splits data property keys into two `DynamicFrames`. The output is a collection of `DynamicFrames`: one with selected data property keys, and one with the remaining data property keys.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the transform node.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The data inputs identified by their node names.
- `Paths` – *Required:* An array of UTF-8 strings.

  A JSON path to a variable in the data structure.

SelectFromCollection structure

Specifies a transform that chooses one `DynamicFrame` from a collection of `DynamicFrames`. The output is the selected `DynamicFrame`

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the transform node.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

The data inputs identified by their node names.

- `Index` – *Required:* Number (integer), not more than None.

  The index for the DynamicFrame to be selected.

FillMissingValues structure

Specifies a transform that locates records in the dataset that have missing values and adds a new field with a value determined by imputation. The input data set is used to train the machine learning model that determines what the missing value should be.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the transform node.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The data inputs identified by their node names.
- `ImputedPath` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  A JSON path to a variable in the data structure for the dataset that is imputed.
- `FilledPath` – UTF-8 string, matching the Custom string pattern #59.

  A JSON path to a variable in the data structure for the dataset that is filled.

Filter structure

Specifies a transform that splits a dataset into two, based on a filter condition.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the transform node.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The data inputs identified by their node names.

- `LogicalOperator` – *Required:* UTF-8 string (valid values: `AND` | `OR`).

  The operator used to filter rows by comparing the key value to a specified value.
- `Filters` – *Required:* An array of FilterExpression objects.

  Specifies a filter expression.

FilterExpression structure

Specifies a filter expression.

Fields

- `Operation` – *Required:* UTF-8 string (valid values: `EQ` | `LT` | `GT` | `LTE` | `GTE` | `REGEX` | `ISNULL`).

  The type of operation to perform in the expression.
- `Negated` – Boolean.

  Whether the expression is to be negated.
- `Values` – *Required:* An array of FilterValue objects.

  A list of filter values.

FilterValue structure

Represents a single entry in the list of values for a `FilterExpression`.

Fields

- `Type` – *Required:* UTF-8 string (valid values: `COLUMNEXTRACTED` | `CONSTANT`).

  The type of filter value.
- `Value` – *Required:* An array of UTF-8 strings.

  The value to be associated.

CustomCode structure

Specifies a transform that uses custom code you provide to perform the data transformation. The output is a collection of DynamicFrames.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the transform node.
- `Inputs` – *Required:* An array of UTF-8 strings, at least 1 string.

  The data inputs identified by their node names.
- `Code` – *Required:* UTF-8 string, matching the Custom string pattern #52.

  The custom code that is used to perform the data transformation.
- `ClassName` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name defined for the custom code node class.
- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for the custom code transform.

SparkSQL structure

Specifies a transform where you enter a SQL query using Spark SQL syntax to transform the data. The output is a single `DynamicFrame`.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the transform node.
- `Inputs` – *Required:* An array of UTF-8 strings, at least 1 string.

  The data inputs identified by their node names. You can associate a table name with each input node to use in the SQL query. The name you choose must meet the Spark SQL naming restrictions.
- `SqlQuery` – *Required:* UTF-8 string, matching the Custom string pattern #60.

A SQL query that must use Spark SQL syntax and return a single data set.

- `SqlAliases` – *Required:* An array of SqlAlias objects.

  A list of aliases. An alias allows you to specify what name to use in the SQL for a given input. For example, you have a datasource named "MyDataSource". If you specify `From` as MyDataSource, and `Alias` as SqlName, then in your SQL you can do:

  ```
  select * from SqlName
  ```

  and that gets data from MyDataSource.
- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for the SparkSQL transform.

SqlAlias structure

Represents a single entry in the list of values for `SqlAliases`.

Fields

- `From` – *Required:* UTF-8 string, matching the Custom string pattern #58.

  A table, or a column in a table.
- `Alias` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  A temporary name given to a table, or a column in a table.

DropNullFields structure

Specifies a transform that removes columns from the dataset if all values in the column are 'null'. By default, AWS Glue Studio will recognize null objects, but some values such as empty strings, strings that are "null", -1 integers or other placeholders such as zeros, are not automatically recognized as nulls.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the transform node.

- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The data inputs identified by their node names.
- `NullCheckBoxList` – A NullCheckBoxList object.

  A structure that represents whether certain values are recognized as null values for removal.
- `NullTextList` – An array of NullValueField objects, not more than 50 structures.

  A structure that specifies a list of NullValueField structures that represent a custom null value such as zero or other value being used as a null placeholder unique to the dataset.

  The `DropNullFields` transform removes custom null values only if both the value of the null placeholder and the datatype match the data.

NullCheckBoxList structure

Represents whether certain values are recognized as null values for removal.

Fields

- `IsEmpty` – Boolean.

  Specifies that an empty string is considered as a null value.
- `IsNullString` – Boolean.

  Specifies that a value spelling out the word 'null' is considered as a null value.
- `IsNegOne` – Boolean.

  Specifies that an integer value of -1 is considered as a null value.

NullValueField structure

Represents a custom null value such as a zeros or other value being used as a null placeholder unique to the dataset.

Fields

- `Value` – *Required:* UTF-8 string, matching the Custom string pattern #59.

The value of the null placeholder.

- `Datatype` – *Required:* A Datatype object.

  The datatype of the value.

Datatype structure

A structure representing the datatype of the value.

Fields

- `Id` – *Required:* UTF-8 string, matching the Custom string pattern #58.

  The datatype of the value.
- `Label` – *Required:* UTF-8 string, matching the Custom string pattern #58.

  A label assigned to the datatype.

Merge structure

Specifies a transform that merges a `DynamicFrame` with a staging `DynamicFrame` based on the specified primary keys to identify records. Duplicate records (records with the same primary keys) are not de-duplicated.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the transform node.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 2 or more than 2 strings.

  The data inputs identified by their node names.
- `Source` – *Required:* UTF-8 string, matching the Custom string pattern #58.

  The source `DynamicFrame` that will be merged with a staging `DynamicFrame`.
- `PrimaryKeys` – *Required:* An array of UTF-8 strings.

  The list of primary key fields to match records from the source and staging dynamic frames.

Union structure

Specifies a transform that combines the rows from two or more datasets into a single result.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the transform node.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 2 or more than 2 strings.

  The node ID inputs to the transform.
- `UnionType` – *Required:* UTF-8 string (valid values: `ALL` | `DISTINCT`).

  Indicates the type of Union transform.

  Specify `ALL` to join all rows from data sources to the resulting DynamicFrame. The resulting union does not remove duplicate rows.

  Specify `DISTINCT` to remove duplicate rows in the resulting DynamicFrame.

PIIDetection structure

Specifies a transform that identifies, removes or masks PII data.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the transform node.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The node ID inputs to the transform.
- `PiiType` – *Required:* UTF-8 string (valid values: `RowAudit` | `RowHashing` | `RowMasking` | `RowPartialMasking` | `ColumnAudit` | `ColumnHashing` | `ColumnMasking`).

  Indicates the type of PIIDetection transform.
- `EntityTypesToDetect` – *Required:* An array of UTF-8 strings.

  Indicates the types of entities the PIIDetection transform will identify as PII data.

PII type entities include: PERSON_NAME, DATE, USA_SNN, EMAIL, USA_ITIN, USA_PASSPORT_NUMBER, PHONE_NUMBER, BANK_ACCOUNT, IP_ADDRESS, MAC_ADDRESS, USA_CPT_CODE, USA_HCPCS_CODE, USA_NATIONAL_DRUG_CODE, USA_MEDICARE_BENEFICIARY_IDENTIFIER, USA_HEALTH_INSURANCE_CLAIM_NUMBER,CREDIT_CARD,USA_NATIONAL_PROVIDER_IDENTIFIER,USA

- `OutputColumnName` – UTF-8 string, matching the Custom string pattern #59.

  Indicates the output column name that will contain any entity type detected in that row.

- `SampleFraction` – Number (double), not more than 1.

  Indicates the fraction of the data to sample when scanning for PII entities.

- `ThresholdFraction` – Number (double), not more than 1.

  Indicates the fraction of the data that must be met in order for a column to be identified as PII data.

- `MaskValue` – UTF-8 string, not more than 256 bytes long, matching the Custom string pattern #56.

  Indicates the value that will replace the detected entity.

- `RedactText` – UTF-8 string, matching the Custom string pattern #59.

  Specifies whether to redact the detected PII text. When set to `true`, PII content is replaced with redaction characters.

- `RedactChar` – UTF-8 string, matching the Custom string pattern #59.

  The character used to replace detected PII content when redaction is enabled. The default redaction character is `*`.

- `MatchPattern` – UTF-8 string, matching the Custom string pattern #59.

  A regular expression pattern used to identify additional PII content beyond the standard detection algorithms.

- `NumLeftCharsToExclude` – Number (integer), not more than None.

  The number of characters to exclude from redaction on the left side of detected PII content. This allows preserving context around the sensitive data.

- `NumRightCharsToExclude` – Number (integer), not more than None.

The number of characters to exclude from redaction on the right side of detected PII content. This allows preserving context around the sensitive data.

- `DetectionParameters` – UTF-8 string, matching the Custom string pattern #59.

  Additional parameters for configuring PII detection behavior and sensitivity settings.
- `DetectionSensitivity` – UTF-8 string, matching the Custom string pattern #59.

  The sensitivity level for PII detection. Higher sensitivity levels detect more potential PII but may result in more false positives.

Aggregate structure

Specifies a transform that groups rows by chosen fields and computes the aggregated value by specified function.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the transform node.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  Specifies the fields and rows to use as inputs for the aggregate transform.
- `Groups` – *Required:* An array of UTF-8 strings.

  Specifies the fields to group by.
- `Aggs` – *Required:* An array of AggregateOperation objects, not less than 1 or more than 30 structures.

  Specifies the aggregate functions to be performed on specified fields.

DropDuplicates structure

Specifies a transform that removes rows of repeating data from a data set.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

The name of the transform node.

- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The data inputs identified by their node names.
- `Columns` – An array of UTF-8 strings.

  The name of the columns to be merged or removed if repeating.

GovernedCatalogTarget structure

Specifies a data target that writes to Amazon S3 using the AWS Glue Data Catalog.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data target.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The nodes that are inputs to the data target.
- `PartitionKeys` – An array of UTF-8 strings.

  Specifies native partitioning using a sequence of keys.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the table in the database to write to.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the database to write to.
- `SchemaChangePolicy` – A CatalogSchemaChangePolicy object.

  A policy that specifies update behavior for the governed catalog.

GovernedCatalogSource structure

Specifies the data store in the governed AWS Glue Data Catalog.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data store.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The database to read from.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The database table to read from.
- `PartitionPredicate` – UTF-8 string, matching the Custom string pattern #59.

  Partitions satisfying this predicate are deleted. Files within the retention period in these partitions are not deleted. Set to `""` – empty by default.
- `AdditionalOptions` – A S3SourceAdditionalOptions object.

  Specifies additional connection options.

AggregateOperation structure

Specifies the set of parameters needed to perform aggregation in the aggregate transform.

Fields

- `Column` – *Required:* An array of UTF-8 strings.

  Specifies the column on the data set on which the aggregation function will be applied.
- `AggFunc` – *Required:* UTF-8 string (valid values: `avg` | `countDistinct` | `count` | `first` | `last` | `kurtosis` | `max` | `min` | `skewness` | `stddev_samp` | `stddev_pop` | `sum` | `sumDistinct` | `var_samp` | `var_pop`).

  Specifies the aggregation function to apply.

  Possible aggregation functions include: avg countDistinct, count, first, last, kurtosis, max, min, skewness, stddev_samp, stddev_pop, sum, sumDistinct, var_samp, var_pop

GlueSchema structure

Specifies a user-defined schema when a schema cannot be determined by AWS Glue.

Fields

- `Columns` – An array of GlueStudioSchemaColumn objects.

  Specifies the column definitions that make up a AWS Glue schema.

GlueStudioSchemaColumn structure

Specifies a single column in a AWS Glue schema definition.

Fields

- `Name` – *Required:* UTF-8 string, not more than 1024 bytes long, matching the Single-line string pattern.

  The name of the column in the AWS Glue Studio schema.
- `Type` – UTF-8 string, not more than 131072 bytes long, matching the Single-line string pattern.

  The hive type for this column in the AWS Glue Studio schema.
- `GlueStudioType` – UTF-8 string, not more than 131072 bytes long, matching the Single-line string pattern.

  The data type of the column as defined in AWS Glue Studio.

GlueStudioColumn structure

Specifies a single column in AWS GlueStudio.

Fields

- `Key` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The key of the column in AWS Glue Studio.
- `FullPath` – *Required:* An array of UTF-8 strings.

  TThe full URL of the column in AWS Glue Studio.

- `Type` – *Required:* UTF-8 string (valid values: `array="ARRAY"` | `bigint="BIGINT"` | `bigint array="BIGINT_ARRAY"` | `binary="BINARY"` | `binary array="BINARY_ARRAY"` | `boolean="BOOLEAN"` | `boolean array="BOOLEAN_ARRAY"` | `byte="BYTE"` | `byte array="BYTE_ARRAY"` | `char="CHAR"` | `char array="CHAR_ARRAY"` | `choice="CHOICE"` | `choice array="CHOICE_ARRAY"` | `date="DATE"` | `date array="DATE_ARRAY"` | `decimal="DECIMAL"` | `decimal array="DECIMAL_ARRAY"` | `double="DOUBLE"` | `double array="DOUBLE_ARRAY"` | `enum="ENUM"` | `enum array="ENUM_ARRAY"` | `float="FLOAT"` | `float array="FLOAT_ARRAY"` | `int="INT"` | `int array="INT_ARRAY"` | `interval="INTERVAL"` | `interval array="INTERVAL_ARRAY"` | `long="LONG"` | `long array="LONG_ARRAY"` | `object="OBJECT"` | `short="SHORT"` | `short array="SHORT_ARRAY"` | `smallint="SMALLINT"` | `smallint array="SMALLINT_ARRAY"` | `string="STRING"` | `string array="STRING_ARRAY"` | `timestamp="TIMESTAMP"` | `timestamp array="TIMESTAMP_ARRAY"` | `tinyint="TINYINT"` | `tinyint array="TINYINT_ARRAY"` | `varchar="VARCHAR"` | `varchar array="VARCHAR_ARRAY"` | `null="NULL"` | `unknown="UNKNOWN"` | `unknown array="UNKNOWN_ARRAY"`).

  TThe type of the column in AWS Glue Studio.

- `Children` – An array of a structures.

  TThe children of the parent column in AWS Glue Studio.

- `GlueStudioType` – UTF-8 string (valid values: `array="ARRAY"` | `bigint="BIGINT"` | `bigint array="BIGINT_ARRAY"` | `binary="BINARY"` | `binary array="BINARY_ARRAY"` | `boolean="BOOLEAN"` | `boolean array="BOOLEAN_ARRAY"` | `byte="BYTE"` | `byte array="BYTE_ARRAY"` | `char="CHAR"` | `char array="CHAR_ARRAY"` | `choice="CHOICE"` | `choice array="CHOICE_ARRAY"` | `date="DATE"` | `date array="DATE_ARRAY"` | `decimal="DECIMAL"` | `decimal array="DECIMAL_ARRAY"` | `double="DOUBLE"` | `double array="DOUBLE_ARRAY"` | `enum="ENUM"` | `enum array="ENUM_ARRAY"` | `float="FLOAT"` | `float array="FLOAT_ARRAY"` | `int="INT"` | `int array="INT_ARRAY"` | `interval="INTERVAL"` | `interval array="INTERVAL_ARRAY"` | `long="LONG"` | `long array="LONG_ARRAY"` | `object="OBJECT"` | `short="SHORT"` | `short array="SHORT_ARRAY"` | `smallint="SMALLINT"` | `smallint array="SMALLINT_ARRAY"` | `string="STRING"` | `string array="STRING_ARRAY"` | `timestamp="TIMESTAMP"` | `timestamp array="TIMESTAMP_ARRAY"` | `tinyint="TINYINT"` | `tinyint array="TINYINT_ARRAY"` | `varchar="VARCHAR"` | `varchar array="VARCHAR_ARRAY"` | `null="NULL"` | `unknown="UNKNOWN"` | `unknown array="UNKNOWN_ARRAY"`).

  The data type of the column as defined in AWS Glue Studio.

DynamicTransform structure

Specifies the set of parameters needed to perform the dynamic transform.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  Specifies the name of the dynamic transform.
- `TransformName` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  Specifies the name of the dynamic transform as it appears in the AWS Glue Studio visual editor.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  Specifies the inputs for the dynamic transform that are required.
- `Parameters` – An array of TransformConfigParameter objects.

  Specifies the parameters of the dynamic transform.
- `FunctionName` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  Specifies the name of the function of the dynamic transform.
- `Path` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  Specifies the path of the dynamic transform source and config files.
- `Version` – UTF-8 string, matching the Custom string pattern #59.

  This field is not used and will be deprecated in future release.
- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for the dynamic transform.

TransformConfigParameter structure

Specifies the parameters in the config file of the dynamic transform.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #59.

Specifies the name of the parameter in the config file of the dynamic transform.

- `Type` – *Required:* UTF-8 string (valid values: `str="STR"` | `int="INT"` | `float="FLOAT"` | `complex="COMPLEX"` | `bool="BOOL"` | `list="LIST"` | `null="NULL"`).

  Specifies the parameter type in the config file of the dynamic transform.
- `ValidationRule` – UTF-8 string, matching the Custom string pattern #59.

  Specifies the validation rule in the config file of the dynamic transform.
- `ValidationMessage` – UTF-8 string, matching the Custom string pattern #59.

  Specifies the validation message in the config file of the dynamic transform.
- `Value` – An array of UTF-8 strings.

  Specifies the value of the parameter in the config file of the dynamic transform.
- `ListType` – UTF-8 string (valid values: `str="STR"` | `int="INT"` | `float="FLOAT"` | `complex="COMPLEX"` | `bool="BOOL"` | `list="LIST"` | `null="NULL"`).

  Specifies the list type of the parameter in the config file of the dynamic transform.
- `IsOptional` – Boolean.

  Specifies whether the parameter is optional or not in the config file of the dynamic transform.

EvaluateDataQuality structure

Specifies your data quality evaluation criteria.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data quality evaluation.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The inputs of your data quality evaluation.
- `Ruleset` – *Required:* UTF-8 string, not less than 1 or more than 65536 bytes long, matching the Custom string pattern #57.

  The ruleset for your data quality evaluation.

- `Output` – UTF-8 string (valid values: `PrimaryInput` | `EvaluationResults`).

  The output of your data quality evaluation.
- `PublishingOptions` – A DQResultsPublishingOptions object.

  Options to configure how your results are published.
- `StopJobOnFailureOptions` – A DQStopJobOnFailureOptions object.

  Options to configure how your job will stop if your data quality evaluation fails.

DQResultsPublishingOptions structure

Options to configure how your data quality evaluation results are published.

Fields

- `EvaluationContext` – UTF-8 string, matching the Custom string pattern #58.

  The context of the evaluation.
- `ResultsS3Prefix` – UTF-8 string, matching the Custom string pattern #59.

  The Amazon S3 prefix prepended to the results.
- `CloudWatchMetricsEnabled` – Boolean.

  Enable metrics for your data quality results.
- `ResultsPublishingEnabled` – Boolean.

  Enable publishing for your data quality results.

DQStopJobOnFailureOptions structure

Options to configure how your job will stop if your data quality evaluation fails.

Fields

- `StopJobOnFailureTiming` – UTF-8 string (valid values: `Immediate` | `AfterDataLoad`).

  When to stop job if your data quality evaluation fails. Options are Immediate or AfterDataLoad.

EvaluateDataQualityMultiFrame structure

Specifies your data quality evaluation criteria.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the data quality evaluation.
- `Inputs` – *Required:* An array of UTF-8 strings, at least 1 string.

  The inputs of your data quality evaluation. The first input in this list is the primary data source.
- `AdditionalDataSources` – A map array of key-value pairs.

  Each key is a UTF-8 string, matching the Custom string pattern #61.

  Each value is a UTF-8 string, matching the Custom string pattern #59.

  The aliases of all data sources except primary.
- `Ruleset` – *Required:* UTF-8 string, not less than 1 or more than 65536 bytes long, matching the Custom string pattern #57.

  The ruleset for your data quality evaluation.
- `PublishingOptions` – A DQResultsPublishingOptions object.

  Options to configure how your results are published.
- `AdditionalOptions` – A map array of key-value pairs.

  Each key is a UTF-8 string (valid values: `performanceTuning.caching="CacheOption"` | `observations.scope="ObservationsOption"` | `compositeRuleEvaluation.method="CompositeOption"`).

  Each value is a UTF-8 string.

  Options to configure runtime behavior of the transform.
- `StopJobOnFailureOptions` – A DQStopJobOnFailureOptions object.

  Options to configure how your job will stop if your data quality evaluation fails.

Recipe structure

A AWS Glue Studio node that uses a AWS Glue DataBrew recipe in AWS Glue jobs.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the AWS Glue Studio node.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The nodes that are inputs to the recipe node, identified by id.
- `RecipeReference` – A RecipeReference object.

  A reference to the DataBrew recipe used by the node.
- `RecipeSteps` – An array of RecipeStep objects.

  Transform steps used in the recipe node.

RecipeReference structure

A reference to a AWS Glue DataBrew recipe.

Fields

- `RecipeArn` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The ARN of the DataBrew recipe.
- `RecipeVersion` – *Required:* UTF-8 string, not less than 1 or more than 16 bytes long.

  The RecipeVersion of the DataBrew recipe.

SnowflakeNodeData structure

Specifies configuration for Snowflake nodes in AWS Glue Studio.

Fields

- `SourceType` – UTF-8 string, matching the Custom string pattern #58.

Specifies how retrieved data is specified. Valid values: `"table"`, `"query"`.

- `Connection` – An Option object.

  Specifies a AWS Glue Data Catalog Connection to a Snowflake endpoint.
- `Schema` – UTF-8 string.

  Specifies a Snowflake database schema for your node to use.
- `Table` – UTF-8 string.

  Specifies a Snowflake table for your node to use.
- `Database` – UTF-8 string.

  Specifies a Snowflake database for your node to use.
- `TempDir` – UTF-8 string, matching the Custom string pattern #59.

  Not currently used.
- `IamRole` – An Option object.

  Not currently used.
- `AdditionalOptions` – A map array of key-value pairs.

  Each key is a UTF-8 string, matching the Custom string pattern #59.

  Each value is a UTF-8 string, matching the Custom string pattern #59.

  Specifies additional options passed to the Snowflake connector. If options are specified elsewhere in this node, this will take precedence.
- `SampleQuery` – UTF-8 string.

  A SQL string used to retrieve data with the `query` sourcetype.
- `PreAction` – UTF-8 string.

  A SQL string run before the Snowflake connector performs its standard actions.
- `PostAction` – UTF-8 string.

  A SQL string run after the Snowflake connector performs its standard actions.
- `Action` – UTF-8 string.

Specifies what action to take when writing to a table with preexisting data. Valid values: `append`, `merge`, `truncate`, `drop`.

- `Upsert` – Boolean.

  Used when Action is `append`. Specifies the resolution behavior when a row already exists. If true, preexisting rows will be updated. If false, those rows will be inserted.
- `MergeAction` – UTF-8 string, matching the Custom string pattern #58.

  Specifies a merge action. Valid values: `simple`, `custom`. If simple, merge behavior is defined by `MergeWhenMatched` and `MergeWhenNotMatched`. If custom, defined by `MergeClause`.
- `MergeWhenMatched` – UTF-8 string, matching the Custom string pattern #58.

  Specifies how to resolve records that match preexisting data when merging. Valid values: `update`, `delete`.
- `MergeWhenNotMatched` – UTF-8 string, matching the Custom string pattern #58.

  Specifies how to process records that do not match preexisting data when merging. Valid values: `insert`, none.
- `MergeClause` – UTF-8 string.

  A SQL statement that specifies a custom merge behavior.
- `StagingTable` – UTF-8 string.

  The name of a staging table used when performing `merge` or upsert `append` actions. Data is written to this table, then moved to `table` by a generated postaction.
- `SelectedColumns` – An array of Option objects.

  Specifies the columns combined to identify a record when detecting matches for merges and upserts. A list of structures with `value`, `label` and `description` keys. Each structure describes a column.
- `AutoPushdown` – Boolean.

  Specifies whether automatic query pushdown is enabled. If pushdown is enabled, then when a query is run on Spark, if part of the query can be "pushed down" to the Snowflake server, it is pushed down. This improves performance of some queries.
- `TableSchema` – An array of Option objects.

Manually defines the target schema for the node. A list of structures with `value` , `label` and `description` keys. Each structure defines a column.

SnowflakeSource structure

Specifies a Snowflake data source.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the Snowflake data source.
- `Data` – *Required:* A SnowflakeNodeData object.

  Configuration for the Snowflake data source.
- `OutputSchemas` – An array of GlueSchema objects.

  Specifies user-defined schemas for your output data.

SnowflakeTarget structure

Specifies a Snowflake target.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the Snowflake target.
- `Data` – *Required:* A SnowflakeNodeData object.

  Specifies the data of the Snowflake target node.
- `Inputs` – An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The nodes that are inputs to the data target.

ConnectorDataSource structure

Specifies a source generated with standard connection options.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of this source node.
- `ConnectionType` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The `connectionType`, as provided to the underlying AWS Glue library. This node type supports the following connection types:
  - `opensearch`
  - `azuresql`
  - `azurecosmos`
  - `bigquery`
  - `saphana`
  - `teradata`
  - `vertica`
- `Data` – *Required:* A map array of key-value pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

  A map specifying connection options for the node. You can find standard connection options for the corresponding connection type in the Connection parameters section of the AWS Glue documentation.
- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for this source.

ConnectorDataTarget structure

Specifies a target generated with standard connection options.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of this target node.

- `ConnectionType` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The `connectionType`, as provided to the underlying AWS Glue library. This node type supports the following connection types:

  - `opensearch`
  - `azuresql`
  - `azurecosmos`
  - `bigquery`
  - `saphana`
  - `teradata`
  - `vertica`
- `Data` – *Required:* A map array of key-value pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

  A map specifying connection options for the node. You can find standard connection options for the corresponding connection type in the Connection parameters section of the AWS Glue documentation.
- `Inputs` – An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The nodes that are inputs to the data target.

RecipeStep structure

A recipe step used in a AWS Glue Studio data preparation recipe node.

Fields

- `Action` – *Required:* A RecipeAction object.

  The transformation action of the recipe step.
- `ConditionExpressions` – An array of ConditionExpression objects.

  The condition expressions for the recipe step.

RecipeAction structure

Actions defined in the AWS Glue Studio data preparation recipe node.

Fields

- `Operation` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #54.

  The operation of the recipe action.
- `Parameters` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #55.

  Each value is a UTF-8 string, not less than 1 or more than 32768 bytes long.

  The parameters of the recipe action.

ConditionExpression structure

Condition expression defined in the AWS Glue Studio data preparation recipe node.

Fields

- `Condition` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #54.

  The condition of the condition expression.
- `Value` – UTF-8 string, not more than 1024 bytes long.

  The value of the condition expression.
- `TargetColumn` – *Required:* UTF-8 string, not less than 1 or more than 1024 bytes long.

  The target column of the condition expressions.

S3CatalogIcebergSource structure

Specifies an Apache Iceberg data source that is registered in the AWS Glue Data Catalog. The Iceberg data source must be stored in Amazon S3.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the Iceberg data source.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the database to read from.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the table in the database to read from.
- `AdditionalIcebergOptions` – A map array of key-value pairs.

  Each key is a UTF-8 string, matching the Custom string pattern #59.

  Each value is a UTF-8 string, matching the Custom string pattern #59.

  Specifies additional connection options for the Iceberg data source.
- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for the Iceberg source.

CatalogIcebergSource structure

Specifies an Apache Iceberg data source that is registered in the AWS Glue Data Catalog.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the Iceberg data source.
- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the database to read from.
- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

The name of the table in the database to read from.

- `AdditionalIcebergOptions` – A map array of key-value pairs.

  Each key is a UTF-8 string, matching the Custom string pattern #59.

  Each value is a UTF-8 string, matching the Custom string pattern #59.

  Specifies additional connection options for the Iceberg data source.

- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for the Iceberg source.

S3IcebergCatalogTarget structure

Specifies an Apache Iceberg catalog target that writes data to Amazon S3 and registers the table in the AWS Glue Data Catalog.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the Iceberg catalog target.

- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The input connection for the Iceberg catalog target.

- `PartitionKeys` – An array of UTF-8 strings.

  A list of partition keys for the Iceberg table.

- `Table` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the table to write to in the catalog.

- `Database` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The name of the database to write to.

- `AdditionalOptions` – A map array of key-value pairs.

  Each key is a UTF-8 string, matching the Custom string pattern #59.

  Each value is a UTF-8 string, matching the Custom string pattern #59.

Specifies additional connection options for the Iceberg catalog target.

- `SchemaChangePolicy` – A CatalogSchemaChangePolicy object.

  The policy for handling schema changes in the catalog target.
- `AutoDataQuality` – An AutoDataQuality object.

  Specifies whether to automatically enable data quality evaluation for the S3 Iceberg catalog target. When set to `true`, data quality checks are performed automatically during the write operation.

DynamoDBELTConnectorSource structure

Specifies a DynamoDB ELT connector source for extracting data from DynamoDB tables.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the DynamoDB ELT connector source.
- `ConnectionOptions` – A DDBELTConnectionOptions object.

  The connection options for the DynamoDB ELT connector source.
- `OutputSchemas` – An array of GlueSchema objects.

  Specifies the data schema for the DynamoDB ELT connector source.

DDBELTConnectionOptions structure

Specifies connection options for DynamoDB ELT (Extract, Load, Transform) operations. This structure contains configuration parameters for connecting to and extracting data from DynamoDB tables using the ELT connector.

Fields

- `DynamodbExport` – UTF-8 string (valid values: ddb | `s3`).

  Specifies the export type for DynamoDB data extraction. This parameter determines how data is exported from the DynamoDB table during the ELT process.

- `DynamodbUnnestDDBJson` – Boolean.

  A boolean value that specifies whether to unnest DynamoDB JSON format during data extraction. When set to `true`, the connector will flatten nested JSON structures from DynamoDB items. When set to `false`, the original DynamoDB JSON structure is preserved.
- `DynamodbTableArn` – *Required:* UTF-8 string, matching the Custom string pattern #59.

  The Amazon Resource Name (ARN) of the DynamoDB table to extract data from. This parameter specifies the source table for the ELT operation.
- `DynamodbS3Bucket` – UTF-8 string, matching the Custom string pattern #59.

  The name of the Amazon S3 bucket used for intermediate storage during the DynamoDB ELT process. This bucket is used to temporarily store exported DynamoDB data before it is processed by the ELT job.
- `DynamodbS3Prefix` – UTF-8 string, matching the Custom string pattern #59.

  The S3 object key prefix for files stored in the intermediate S3 bucket during the DynamoDB ELT process. This prefix helps organize and identify the temporary files created during data extraction.
- `DynamodbS3BucketOwner` – UTF-8 string, matching the Custom string pattern #59.

  The AWS account ID of the owner of the S3 bucket specified in `DynamodbS3Bucket`. This parameter is required when the S3 bucket is owned by a different AWS account than the one running the ELT job, enabling cross-account access to the intermediate storage bucket.
- `DynamodbStsRoleArn` – UTF-8 string, matching the Custom string pattern #59.

  The Amazon Resource Name (ARN) of the AWS Security Token Service (STS) role to assume for accessing DynamoDB and S3 resources during the ELT operation. This role must have the necessary permissions to read from the DynamoDB table and write to the intermediate S3 bucket.

DDBELTCatalogAdditionalOptions structure

Specifies additional options for DynamoDB ELT catalog operations.

Fields

- `DynamodbExport` – UTF-8 string, matching the Custom string pattern #59.

Specifies the DynamoDB export configuration for the ELT operation.

- `DynamodbUnnestDDBJson` – Boolean.

  Specifies whether to unnest DynamoDB JSON format. When set to `true`, nested JSON structures in DynamoDB items are flattened.

Route structure

Specifies a route node that directs data to different output paths based on defined filtering conditions.

Fields

- `Name` – *Required:* UTF-8 string, matching the Custom string pattern #61.

  The name of the route node.
- `Inputs` – *Required:* An array of UTF-8 strings, not less than 1 or more than 1 strings.

  The input connection for the route node.
- `GroupFiltersList` – *Required:* An array of GroupFilters objects.

  A list of group filters that define the routing conditions and criteria for directing data to different output paths.

GroupFilters structure

Specifies a group of filters with a logical operator that determines how the filters are combined to evaluate routing conditions.

Fields

- `GroupName` – *Required:* UTF-8 string, matching the Custom string pattern #58.

  The name of the filter group.
- `Filters` – *Required:* An array of FilterExpression objects.

  A list of filter expressions that define the conditions for this group.
- `LogicalOperator` – *Required:* UTF-8 string (valid values: `AND` | `OR`).

The logical operator used to combine the filters in this group. Determines whether all filters must match (AND) or any filter can match (OR).

AutoDataQuality structure

Specifies configuration options for automatic data quality evaluation in AWS Glue jobs. This structure enables automated data quality checks and monitoring during ETL operations, helping to ensure data integrity and reliability without manual intervention.

Fields

- `IsEnabled` – Boolean.

  Specifies whether automatic data quality evaluation is enabled. When set to `true`, data quality checks are performed automatically.
- `EvaluationContext` – UTF-8 string, matching the Custom string pattern #59.

  The evaluation context for the automatic data quality checks. This defines the scope and parameters for the data quality evaluation.

Jobs API

The Jobs API describes jobs data types and contains APIs for working with jobs, job runs, and triggers in AWS Glue.

Topics

- Jobs
- Job runs
- Triggers

Jobs

The Jobs API describes the data types and API related to creating, updating, deleting, or viewing jobs in AWS Glue.

Data types

- Job structure
- ExecutionProperty structure
- NotificationProperty structure
- JobCommand structure
- ConnectionsList structure
- JobUpdate structure
- SourceControlDetails structure

Job structure

Specifies a job definition.

Fields

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name you assign to this job definition.
- `JobMode` – UTF-8 string (valid values: `SCRIPT=""` | `VISUAL=""` | `NOTEBOOK=""`).

  A mode that describes how a job was created. Valid values are:
  - `SCRIPT` - The job was created using the AWS Glue Studio script editor.
  - `VISUAL` - The job was created using the AWS Glue Studio visual editor.
  - `NOTEBOOK` - The job was created using an interactive sessions notebook.

  When the `JobMode` field is missing or null, `SCRIPT` is assigned as the default value.
- `JobRunQueuingEnabled` – Boolean.

  Specifies whether job run queuing is enabled for the job runs for this job.

  A value of true means job run queuing is enabled for the job runs. If false or not populated, the job runs will not be considered for queueing.

  If this field does not match the value set in the job run, then the value from the job run field will be used.

- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the job.

- `LogUri` – UTF-8 string.

  This field is reserved for future use.

- `Role` – UTF-8 string.

  The name or Amazon Resource Name (ARN) of the IAM role associated with this job.

- `CreatedOn` – Timestamp.

  The time and date that this job definition was created.

- `LastModifiedOn` – Timestamp.

  The last point in time when this job definition was modified.

- `ExecutionProperty` – An ExecutionProperty object.

  An `ExecutionProperty` specifying the maximum number of concurrent runs allowed for this job.

- `Command` – A JobCommand object.

  The `JobCommand` that runs this job.

- `DefaultArguments` – A map array of key-value pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

  The default arguments for every run of this job, specified as name-value pairs.

  You can specify arguments here that your own job-execution script consumes, as well as arguments that AWS Glue itself consumes.

  Job arguments may be logged. Do not pass plaintext secrets as arguments. Retrieve secrets from a AWS Glue Connection, AWS Secrets Manager or other secret management mechanism if you intend to keep them within the Job.

For information about how to specify and consume your own Job arguments, see the Calling AWS Glue APIs in Python topic in the developer guide.

For information about the arguments you can provide to this field when configuring Spark jobs, see the Special Parameters Used by AWS Glue topic in the developer guide.

For information about the arguments you can provide to this field when configuring Ray jobs, see Using job parameters in Ray jobs in the developer guide.

- `NonOverridableArguments` – A map array of key-value pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

  Arguments for this job that are not overridden when providing job arguments in a job run, specified as name-value pairs.
- `Connections` – A ConnectionsList object.

  The connections used for this job.
- `MaxRetries` – Number (integer).

  The maximum number of times to retry this job after a JobRun fails.
- `AllocatedCapacity` – Number (integer).

  This field is deprecated. Use `MaxCapacity` instead.

  The number of AWS Glue data processing units (DPUs) allocated to runs of this job. You can allocate a minimum of 2 DPUs; the default is 10. A DPU is a relative measure of processing power that consists of 4 vCPUs of compute capacity and 16 GB of memory. For more information, see the AWS Glue pricing page.
- `Timeout` – Number (integer), at least 1.

  The job timeout in minutes. This is the maximum time that a job run can consume resources before it is terminated and enters `TIMEOUT` status.

  Jobs must have timeout values less than 7 days or 10080 minutes. Otherwise, the jobs will throw an exception.

When the value is left blank, the timeout is defaulted to 2880 minutes.

Any existing AWS Glue jobs that had a timeout value greater than 7 days will be defaulted to 7 days. For instance if you have specified a timeout of 20 days for a batch job, it will be stopped on the 7th day.

For streaming jobs, if you have set up a maintenance window, it will be restarted during the maintenance window after 7 days.

- `MaxCapacity` – Number (double).

  For Glue version 1.0 or earlier jobs, using the standard worker type, the number of AWS Glue data processing units (DPUs) that can be allocated when this job runs. A DPU is a relative measure of processing power that consists of 4 vCPUs of compute capacity and 16 GB of memory. For more information, see the AWS Glue pricing page.

  For Glue version 2.0 or later jobs, you cannot specify a `Maximum capacity`. Instead, you should specify a `Worker type` and the `Number of workers`.

  Do not set `MaxCapacity` if using `WorkerType` and `NumberOfWorkers`.

  The value that can be allocated for `MaxCapacity` depends on whether you are running a Python shell job, an Apache Spark ETL job, or an Apache Spark streaming ETL job:

  - When you specify a Python shell job (`JobCommand.Name`="pythonshell"), you can allocate either 0.0625 or 1 DPU. The default is 0.0625 DPU.
  - When you specify an Apache Spark ETL job (`JobCommand.Name`="glueetl") or Apache Spark streaming ETL job (`JobCommand.Name`="gluestreaming"), you can allocate from 2 to 100 DPUs. The default is 10 DPUs. This job type cannot have a fractional DPU allocation.

- `WorkerType` – UTF-8 string (valid values: `Standard=""` | `G.1X=""` | `G.2X=""` | `G.025X=""` | `G.4X=""` | `G.8X=""` | `Z.2X=""`).

  The type of predefined worker that is allocated when a job runs.

  AWS Glue provides multiple worker types to accommodate different workload requirements:

  G Worker Types (General-purpose compute workers):

  - G.1X: 1 DPU (4 vCPUs, 16 GB memory, 94GB disk)
  - G.2X: 2 DPU (8 vCPUs, 32 GB memory, 138GB disk)

  - G.4X: 4 DPU (16 vCPUs, 64 GB memory, 256GB disk)
  - G.8X: 8 DPU (32 vCPUs, 128 GB memory, 512GB disk)
  - G.12X: 12 DPU (48 vCPUs, 192 GB memory, 768GB disk)
  - G.16X: 16 DPU (64 vCPUs, 256 GB memory, 1024GB disk)

  R Worker Types (Memory-optimized workers):

  - R.1X: 1 M-DPU (4 vCPUs, 32 GB memory)
  - R.2X: 2 M-DPU (8 vCPUs, 64 GB memory)
  - R.4X: 4 M-DPU (16 vCPUs, 128 GB memory)
  - R.8X: 8 M-DPU (32 vCPUs, 256 GB memory)

- `NumberOfWorkers` – Number (integer).

  The number of workers of a defined `workerType` that are allocated when a job runs.

- `SecurityConfiguration` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the `SecurityConfiguration` structure to be used with this job.

- `NotificationProperty` – A NotificationProperty object.

  Specifies configuration properties of a job notification.

- `Running` – Boolean.

  This field is reserved for future use.

- `GlueVersion` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #47.

  In Spark jobs, `GlueVersion` determines the versions of Apache Spark and Python that AWS Glue available in a job. The Python version indicates the version supported for jobs of type Spark.

  Ray jobs should set `GlueVersion` to `4.0` or greater. However, the versions of Ray, Python and additional libraries available in your Ray job are determined by the `Runtime` parameter of the Job command.

  For more information about the available AWS Glue versions and corresponding Spark and Python versions, see Glue version in the developer guide.

Jobs that are created without specifying a Glue version default to Glue 0.9.

- `CodeGenConfigurationNodes` – A map array of key-value pairs.

  Each key is a UTF-8 string, matching the Custom string pattern #58.

  Each value is a A CodeGenConfigurationNode object.

  The representation of a directed acyclic graph on which both the Glue Studio visual component and Glue Studio code generation is based.
- `ExecutionClass` – UTF-8 string, not more than 16 bytes long (valid values: `FLEX=""` | `STANDARD=""`).

  Indicates whether the job is run with a standard or flexible execution class. The standard execution class is ideal for time-sensitive workloads that require fast job startup and dedicated resources.

  The flexible execution class is appropriate for time-insensitive jobs whose start and completion times may vary.

  Only jobs with AWS Glue version 3.0 and above and command type `glueetl` will be allowed to set `ExecutionClass` to `FLEX`. The flexible execution class is available for Spark jobs.
- `SourceControlDetails` – A SourceControlDetails object.

  The details for a source control configuration for a job, allowing synchronization of job artifacts to or from a remote repository.
- `MaintenanceWindow` – UTF-8 string, matching the Custom string pattern #34.

  This field specifies a day of the week and hour for a maintenance window for streaming jobs. AWS Glue periodically performs maintenance activities. During these maintenance windows, AWS Glue will need to restart your streaming jobs.

  AWS Glue will restart the job within 3 hours of the specified maintenance window. For instance, if you set up the maintenance window for Monday at 10:00AM GMT, your jobs will be restarted between 10:00AM GMT to 1:00PM GMT.
- `ProfileName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of an AWS Glue usage profile associated with the job.

ExecutionProperty structure

An execution property of a job.

Fields

- `MaxConcurrentRuns` – Number (integer).

  The maximum number of concurrent runs allowed for the job. The default is 1. An error is returned when this threshold is reached. The maximum value you can specify is controlled by a service limit.

NotificationProperty structure

Specifies configuration properties of a notification.

Fields

- `NotifyDelayAfter` – Number (integer), at least 1.

  After a job run starts, the number of minutes to wait before sending a job run delay notification.

JobCommand structure

Specifies code that runs when a job is run.

Fields

- `Name` – UTF-8 string.

  The name of the job command. For an Apache Spark ETL job, this must be `glueetl`. For a Python shell job, it must be `pythonshell`. For an Apache Spark streaming ETL job, this must be `gluestreaming`. For a Ray job, this must be `glueray`.
- `ScriptLocation` – UTF-8 string, not more than 400000 bytes long.

  Specifies the Amazon Simple Storage Service (Amazon S3) path to a script that runs a job.
- `PythonVersion` – UTF-8 string, matching the Custom string pattern #48.

  The Python version being used to run a Python shell job. Allowed values are 2 or 3.
- `Runtime` – UTF-8 string, not more than 64 bytes long, matching the Custom string pattern #33.

In Ray jobs, Runtime is used to specify the versions of Ray, Python and additional libraries available in your environment. This field is not used in other job types. For supported runtime environment values, see [Supported Ray runtime environments]() in the AWS Glue Developer Guide.

ConnectionsList structure

Specifies the connections used by a job.

Fields

- `Connections` – An array of UTF-8 strings, not more than 1000 strings.

  A list of connections used by the job.

JobUpdate structure

Specifies information used to update an existing job definition. The previous job definition is completely overwritten by this information.

Fields

- `JobMode` – UTF-8 string (valid values: `SCRIPT=""` | `VISUAL=""` | `NOTEBOOK=""`).

  A mode that describes how a job was created. Valid values are:

  - `SCRIPT` - The job was created using the AWS Glue Studio script editor.
  - `VISUAL` - The job was created using the AWS Glue Studio visual editor.
  - `NOTEBOOK` - The job was created using an interactive sessions notebook.

  When the `JobMode` field is missing or null, `SCRIPT` is assigned as the default value.
- `JobRunQueuingEnabled` – Boolean.

  Specifies whether job run queuing is enabled for the job runs for this job.

  A value of true means job run queuing is enabled for the job runs. If false or not populated, the job runs will not be considered for queueing.

  If this field does not match the value set in the job run, then the value from the job run field will be used.

- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  Description of the job being defined.
- `LogUri` – UTF-8 string.

  This field is reserved for future use.
- `Role` – UTF-8 string.

  The name or Amazon Resource Name (ARN) of the IAM role associated with this job (required).
- `ExecutionProperty` – An ExecutionProperty object.

  An `ExecutionProperty` specifying the maximum number of concurrent runs allowed for this job.
- `Command` – A JobCommand object.

  The `JobCommand` that runs this job (required).
- `DefaultArguments` – A map array of key-value pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

  The default arguments for every run of this job, specified as name-value pairs.

  You can specify arguments here that your own job-execution script consumes, as well as arguments that AWS Glue itself consumes.

  Job arguments may be logged. Do not pass plaintext secrets as arguments. Retrieve secrets from a AWS Glue Connection, AWS Secrets Manager or other secret management mechanism if you intend to keep them within the Job.

  For information about how to specify and consume your own Job arguments, see the Calling AWS Glue APIs in Python topic in the developer guide.

  For information about the arguments you can provide to this field when configuring Spark jobs, see the Special Parameters Used by AWS Glue topic in the developer guide.

  For information about the arguments you can provide to this field when configuring Ray jobs, see Using job parameters in Ray jobs in the developer guide.

- `NonOverridableArguments` – A map array of key-value pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

  Arguments for this job that are not overridden when providing job arguments in a job run, specified as name-value pairs.
- `Connections` – A ConnectionsList object.

  The connections used for this job.
- `MaxRetries` – Number (integer).

  The maximum number of times to retry this job if it fails.
- `AllocatedCapacity` – Number (integer).

  This field is deprecated. Use `MaxCapacity` instead.

  The number of AWS Glue data processing units (DPUs) to allocate to this job. You can allocate a minimum of 2 DPUs; the default is 10. A DPU is a relative measure of processing power that consists of 4 vCPUs of compute capacity and 16 GB of memory. For more information, see the AWS Glue pricing page.
- `Timeout` – Number (integer), at least 1.

  The job timeout in minutes. This is the maximum time that a job run can consume resources before it is terminated and enters `TIMEOUT` status.

  Jobs must have timeout values less than 7 days or 10080 minutes. Otherwise, the jobs will throw an exception.

  When the value is left blank, the timeout is defaulted to 2880 minutes.

  Any existing AWS Glue jobs that had a timeout value greater than 7 days will be defaulted to 7 days. For instance if you have specified a timeout of 20 days for a batch job, it will be stopped on the 7th day.

  For streaming jobs, if you have set up a maintenance window, it will be restarted during the maintenance window after 7 days.
- `MaxCapacity` – Number (double).

For Glue version 1.0 or earlier jobs, using the standard worker type, the number of AWS Glue data processing units (DPUs) that can be allocated when this job runs. A DPU is a relative measure of processing power that consists of 4 vCPUs of compute capacity and 16 GB of memory. For more information, see the AWS Glue pricing page.

For Glue version 2.0+ jobs, you cannot specify a `Maximum capacity`. Instead, you should specify a `Worker type` and the `Number of workers`.

Do not set `MaxCapacity` if using `WorkerType` and `NumberOfWorkers`.

The value that can be allocated for `MaxCapacity` depends on whether you are running a Python shell job, an Apache Spark ETL job, or an Apache Spark streaming ETL job:

- When you specify a Python shell job (`JobCommand.Name`="pythonshell"), you can allocate either 0.0625 or 1 DPU. The default is 0.0625 DPU.
- When you specify an Apache Spark ETL job (`JobCommand.Name`="glueetl") or Apache Spark streaming ETL job (`JobCommand.Name`="gluestreaming"), you can allocate from 2 to 100 DPUs. The default is 10 DPUs. This job type cannot have a fractional DPU allocation.

- `WorkerType` – UTF-8 string (valid values: `Standard=""` | `G.1X=""` | `G.2X=""` | `G.025X=""` | `G.4X=""` | `G.8X=""` | `Z.2X=""`).

  The type of predefined worker that is allocated when a job runs. Accepts a value of G.1X, G.2X, G.4X, G.8X or G.025X for Spark jobs. Accepts the value Z.2X for Ray jobs. For more information, see Defining job properties for Spark jobs

- `NumberOfWorkers` – Number (integer).

  The number of workers of a defined `workerType` that are allocated when a job runs.

- `SecurityConfiguration` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the `SecurityConfiguration` structure to be used with this job.

- `NotificationProperty` – A NotificationProperty object.

  Specifies the configuration properties of a job notification.

- `GlueVersion` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #47.

In Spark jobs, `GlueVersion` determines the versions of Apache Spark and Python that AWS Glue available in a job. The Python version indicates the version supported for jobs of type Spark.

Ray jobs should set `GlueVersion` to `4.0` or greater. However, the versions of Ray, Python and additional libraries available in your Ray job are determined by the `Runtime` parameter of the Job command.

For more information about the available AWS Glue versions and corresponding Spark and Python versions, see Glue version in the developer guide.

Jobs that are created without specifying a Glue version default to Glue 0.9.

- `CodeGenConfigurationNodes` – A map array of key-value pairs.

  Each key is a UTF-8 string, matching the Custom string pattern #58.

  Each value is a A CodeGenConfigurationNode object.

  The representation of a directed acyclic graph on which both the Glue Studio visual component and Glue Studio code generation is based.
- `ExecutionClass` – UTF-8 string, not more than 16 bytes long (valid values: `FLEX=""` | `STANDARD=""`).

  Indicates whether the job is run with a standard or flexible execution class. The standard execution-class is ideal for time-sensitive workloads that require fast job startup and dedicated resources.

  The flexible execution class is appropriate for time-insensitive jobs whose start and completion times may vary.

  Only jobs with AWS Glue version 3.0 and above and command type `glueetl` will be allowed to set `ExecutionClass` to `FLEX`. The flexible execution class is available for Spark jobs.
- `SourceControlDetails` – A SourceControlDetails object.

  The details for a source control configuration for a job, allowing synchronization of job artifacts to or from a remote repository.
- `MaintenanceWindow` – UTF-8 string, matching the Custom string pattern #34.

This field specifies a day of the week and hour for a maintenance window for streaming jobs. AWS Glue periodically performs maintenance activities. During these maintenance windows, AWS Glue will need to restart your streaming jobs.

AWS Glue will restart the job within 3 hours of the specified maintenance window. For instance, if you set up the maintenance window for Monday at 10:00AM GMT, your jobs will be restarted between 10:00AM GMT to 1:00PM GMT.

- `ProfileName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of an AWS Glue usage profile associated with the job.

SourceControlDetails structure

The details for a source control configuration for a job, allowing synchronization of job artifacts to or from a remote repository.

Fields

- `Provider` – UTF-8 string (valid values: `GITHUB` | `AWS_CODE_COMMIT`).

  The provider for the remote repository.
- `Repository` – UTF-8 string, not less than 1 or more than 512 bytes long.

  The name of the remote repository that contains the job artifacts.
- `Owner` – UTF-8 string, not less than 1 or more than 512 bytes long.

  The owner of the remote repository that contains the job artifacts.
- `Branch` – UTF-8 string, not less than 1 or more than 512 bytes long.

  An optional branch in the remote repository.
- `Folder` – UTF-8 string, not less than 1 or more than 512 bytes long.

  An optional folder in the remote repository.
- `LastCommitId` – UTF-8 string, not less than 1 or more than 512 bytes long.

  The last commit ID for a commit in the remote repository.
- `LastSyncTimestamp` – UTF-8 string, not less than 1 or more than 512 bytes long.

The date and time that the last job synchronization was performed.

- `AuthStrategy` – UTF-8 string (valid values: `PERSONAL_ACCESS_TOKEN` | `AWS_SECRETS_MANAGER`).

  The type of authentication, which can be an authentication token stored in AWS Secrets Manager, or a personal access token.
- `AuthToken` – UTF-8 string, not less than 1 or more than 512 bytes long.

  The value of an authorization token.

Operations

- CreateJob action (Python: create_job)
- UpdateJob action (Python: update_job)
- GetJob action (Python: get_job)
- GetJobs action (Python: get_jobs)
- DeleteJob action (Python: delete_job)
- ListJobs action (Python: list_jobs)
- BatchGetJobs action (Python: batch_get_jobs)

CreateJob action (Python: create_job)

Creates a new job definition.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name you assign to this job definition. It must be unique in your account.
- `JobMode` – UTF-8 string (valid values: `SCRIPT=""` | `VISUAL=""` | `NOTEBOOK=""`).

  A mode that describes how a job was created. Valid values are:
  - `SCRIPT` - The job was created using the AWS Glue Studio script editor.
  - `VISUAL` - The job was created using the AWS Glue Studio visual editor.

  - `NOTEBOOK` - The job was created using an interactive sessions notebook.

  When the `JobMode` field is missing or null, `SCRIPT` is assigned as the default value.

- `JobRunQueuingEnabled` – Boolean.

  Specifies whether job run queuing is enabled for the job runs for this job.

  A value of true means job run queuing is enabled for the job runs. If false or not populated, the job runs will not be considered for queueing.

  If this field does not match the value set in the job run, then the value from the job run field will be used.

- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  Description of the job being defined.

- `LogUri` – UTF-8 string.

  This field is reserved for future use.

- `Role` – *Required:* UTF-8 string.

  The name or Amazon Resource Name (ARN) of the IAM role associated with this job.

- `ExecutionProperty` – An ExecutionProperty object.

  An `ExecutionProperty` specifying the maximum number of concurrent runs allowed for this job.

- `Command` – *Required:* A JobCommand object.

  The `JobCommand` that runs this job.

- `DefaultArguments` – A map array of key-value pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

  The default arguments for every run of this job, specified as name-value pairs.

  You can specify arguments here that your own job-execution script consumes, as well as arguments that AWS Glue itself consumes.

Job arguments may be logged. Do not pass plaintext secrets as arguments. Retrieve secrets from a AWS Glue Connection, AWS Secrets Manager or other secret management mechanism if you intend to keep them within the Job.

For information about how to specify and consume your own Job arguments, see the Calling AWS Glue APIs in Python topic in the developer guide.

For information about the arguments you can provide to this field when configuring Spark jobs, see the Special Parameters Used by AWS Glue topic in the developer guide.

For information about the arguments you can provide to this field when configuring Ray jobs, see Using job parameters in Ray jobs in the developer guide.

- `NonOverridableArguments` – A map array of key-value pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

  Arguments for this job that are not overridden when providing job arguments in a job run, specified as name-value pairs.
- `Connections` – A ConnectionsList object.

  The connections used for this job.
- `MaxRetries` – Number (integer).

  The maximum number of times to retry this job if it fails.
- `AllocatedCapacity` – Number (integer).

  This parameter is deprecated. Use `MaxCapacity` instead.

  The number of AWS Glue data processing units (DPUs) to allocate to this Job. You can allocate a minimum of 2 DPUs; the default is 10. A DPU is a relative measure of processing power that consists of 4 vCPUs of compute capacity and 16 GB of memory. For more information, see the AWS Glue pricing page.
- `Timeout` – Number (integer), at least 1.

  The job timeout in minutes. This is the maximum time that a job run can consume resources before it is terminated and enters `TIMEOUT` status.

Jobs must have timeout values less than 7 days or 10080 minutes. Otherwise, the jobs will throw an exception.

When the value is left blank, the timeout is defaulted to 2880 minutes.

Any existing AWS Glue jobs that had a timeout value greater than 7 days will be defaulted to 7 days. For instance if you have specified a timeout of 20 days for a batch job, it will be stopped on the 7th day.

For streaming jobs, if you have set up a maintenance window, it will be restarted during the maintenance window after 7 days.

- `MaxCapacity` – Number (double).

  For Glue version 1.0 or earlier jobs, using the standard worker type, the number of AWS Glue data processing units (DPUs) that can be allocated when this job runs. A DPU is a relative measure of processing power that consists of 4 vCPUs of compute capacity and 16 GB of memory. For more information, see the AWS Glue pricing page.

  For Glue version 2.0+ jobs, you cannot specify a `Maximum capacity`. Instead, you should specify a `Worker type` and the `Number of workers`.

  Do not set `MaxCapacity` if using `WorkerType` and `NumberOfWorkers`.

  The value that can be allocated for `MaxCapacity` depends on whether you are running a Python shell job, an Apache Spark ETL job, or an Apache Spark streaming ETL job:

  - When you specify a Python shell job (`JobCommand.Name`="pythonshell"), you can allocate either 0.0625 or 1 DPU. The default is 0.0625 DPU.
  - When you specify an Apache Spark ETL job (`JobCommand.Name`="glueetl") or Apache Spark streaming ETL job (`JobCommand.Name`="gluestreaming"), you can allocate from 2 to 100 DPUs. The default is 10 DPUs. This job type cannot have a fractional DPU allocation.
- `SecurityConfiguration` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the `SecurityConfiguration` structure to be used with this job.
- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

Each value is a UTF-8 string, not more than 256 bytes long.

The tags to use with this job. You may use tags to limit access to the job. For more information about tags in AWS Glue, see AWS Tags in AWS Glue in the developer guide.

- `NotificationProperty` – A NotificationProperty object.

  Specifies configuration properties of a job notification.

- `GlueVersion` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #47.

  In Spark jobs, `GlueVersion` determines the versions of Apache Spark and Python that AWS Glue available in a job. The Python version indicates the version supported for jobs of type Spark.

  Ray jobs should set `GlueVersion` to `4.0` or greater. However, the versions of Ray, Python and additional libraries available in your Ray job are determined by the `Runtime` parameter of the Job command.

  For more information about the available AWS Glue versions and corresponding Spark and Python versions, see Glue version in the developer guide.

  Jobs that are created without specifying a Glue version default to Glue 0.9.

- `NumberOfWorkers` – Number (integer).

  The number of workers of a defined `workerType` that are allocated when a job runs.

- `WorkerType` – UTF-8 string (valid values: `Standard=""` | `G.1X=""` | `G.2X=""` | `G.025X=""` | `G.4X=""` | `G.8X=""` | `Z.2X=""`).

  The type of predefined worker that is allocated when a job runs. Accepts a value of G.1X, G.2X, G.4X, G.8X or G.025X for Spark jobs. Accepts the value Z.2X for Ray jobs.

  - For the `G.1X` worker type, each worker maps to 1 DPU (4 vCPUs, 16 GB of memory) with 94GB disk, and provides 1 executor per worker. We recommend this worker type for workloads such as data transforms, joins, and queries, to offers a scalable and cost effective way to run most jobs.
  - For the `G.2X` worker type, each worker maps to 2 DPU (8 vCPUs, 32 GB of memory) with 138GB disk, and provides 1 executor per worker. We recommend this worker type for

  workloads such as data transforms, joins, and queries, to offers a scalable and cost effective way to run most jobs.
  - For the `G.4X` worker type, each worker maps to 4 DPU (16 vCPUs, 64 GB of memory) with 256GB disk, and provides 1 executor per worker. We recommend this worker type for jobs whose workloads contain your most demanding transforms, aggregations, joins, and queries. This worker type is available only for AWS Glue version 3.0 or later Spark ETL jobs in the following AWS Regions: US East (Ohio), US East (N. Virginia), US West (N. California), US West (Oregon), Asia Pacific (Mumbai), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Canada (Central), Europe (Frankfurt), Europe (Ireland), Europe (London), Europe (Spain), Europe (Stockholm), and South America (São Paulo).
  - For the `G.8X` worker type, each worker maps to 8 DPU (32 vCPUs, 128 GB of memory) with 512GB disk, and provides 1 executor per worker. We recommend this worker type for jobs whose workloads contain your most demanding transforms, aggregations, joins, and queries. This worker type is available only for AWS Glue version 3.0 or later Spark ETL jobs, in the same AWS Regions as supported for the `G.4X` worker type.
  - For the `G.025X` worker type, each worker maps to 0.25 DPU (2 vCPUs, 4 GB of memory) with 84GB disk, and provides 1 executor per worker. We recommend this worker type for low volume streaming jobs. This worker type is only available for AWS Glue version 3.0 or later streaming jobs.
  - For the `Z.2X` worker type, each worker maps to 2 M-DPU (8vCPUs, 64 GB of memory) with 128 GB disk, and provides up to 8 Ray workers based on the autoscaler.
- `CodeGenConfigurationNodes` – A map array of key-value pairs.

  Each key is a UTF-8 string, matching the Custom string pattern #58.

  Each value is a A CodeGenConfigurationNode object.

  The representation of a directed acyclic graph on which both the Glue Studio visual component and Glue Studio code generation is based.
- `ExecutionClass` – UTF-8 string, not more than 16 bytes long (valid values: `FLEX=""` | `STANDARD=""`).

  Indicates whether the job is run with a standard or flexible execution class. The standard execution-class is ideal for time-sensitive workloads that require fast job startup and dedicated resources.

The flexible execution class is appropriate for time-insensitive jobs whose start and completion times may vary.

Only jobs with AWS Glue version 3.0 and above and command type `glueetl` will be allowed to set `ExecutionClass` to `FLEX`. The flexible execution class is available for Spark jobs.

- `SourceControlDetails` – A SourceControlDetails object.

  The details for a source control configuration for a job, allowing synchronization of job artifacts to or from a remote repository.
- `MaintenanceWindow` – UTF-8 string, matching the Custom string pattern #34.

  This field specifies a day of the week and hour for a maintenance window for streaming jobs. AWS Glue periodically performs maintenance activities. During these maintenance windows, AWS Glue will need to restart your streaming jobs.

  AWS Glue will restart the job within 3 hours of the specified maintenance window. For instance, if you set up the maintenance window for Monday at 10:00AM GMT, your jobs will be restarted between 10:00AM GMT to 1:00PM GMT.
- `ProfileName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of an AWS Glue usage profile associated with the job.

Response

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique name that was provided for this job definition.

Errors

- `InvalidInputException`
- `IdempotentParameterMismatchException`
- `AlreadyExistsException`
- `InternalServiceException`

- `OperationTimeoutException`
- `ResourceNumberLimitExceededException`
- `ConcurrentModificationException`

UpdateJob action (Python: update_job)

Updates an existing job definition. The previous job definition is completely overwritten by this information.

Request

- `JobName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the job definition to update.
- `JobUpdate` – *Required:* A JobUpdate object.

  Specifies the values with which to update the job definition. Unspecified configuration is removed or reset to default values.
- `ProfileName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of an AWS Glue usage profile associated with the job.

Response

- `JobName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Returns the name of the updated job definition.

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`

- `ConcurrentModificationException`

GetJob action (Python: get_job)

Retrieves an existing job definition.

Request

- `JobName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the job definition to retrieve.

Response

- `Job` – A Job object.

  The requested job definition.

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`

GetJobs action (Python: get_jobs)

Retrieves all current job definitions.

Request

- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation call.
- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum size of the response.

Response

- `Jobs` – An array of Job objects.

  A list of job definitions.
- `NextToken` – UTF-8 string.

  A continuation token, if not all job definitions have yet been returned.

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`

DeleteJob action (Python: delete_job)

Deletes a specified job definition. If the job definition is not found, no exception is thrown.

Request

- `JobName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the job definition to delete.

Response

- `JobName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the job definition that was deleted.

Errors

- `InvalidInputException`
- `InternalServiceException`

- `OperationTimeoutException`

ListJobs action (Python: list_jobs)

Retrieves the names of all job resources in this AWS account, or the resources with the specified tag. This operation allows you to see which resources are available in your account, and their names.

This operation takes the optional `Tags` field, which you can use as a filter on the response so that tagged resources can be retrieved as a group. If you choose to use tags filtering, only resources with the tag are retrieved.

Request

- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation request.
- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum size of a list to return.
- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  Specifies to return only these tagged resources.

Response

- `JobNames` – An array of UTF-8 strings.

  The names of all jobs in the account, or the jobs with the specified tags.
- `NextToken` – UTF-8 string.

  A continuation token, if the returned list does not contain the last metric available.

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`

BatchGetJobs action (Python: batch_get_jobs)

Returns a list of resource metadata for a given list of job names. After calling the `ListJobs` operation, you can call this operation to access the data to which you have been granted permissions. This operation supports all IAM permissions, including permission conditions that uses tags.

Request

- `JobNames` – *Required:* An array of UTF-8 strings.

  A list of job names, which might be the names returned from the `ListJobs` operation.

Response

- `Jobs` – An array of Job objects.

  A list of job definitions.
- `JobsNotFound` – An array of UTF-8 strings.

  A list of names of jobs not found.

Errors

- `InternalServiceException`
- `OperationTimeoutException`
- `InvalidInputException`

Job runs

The Jobs Runs API describes the data types and API related to starting, stopping, or viewing job runs, and resetting job bookmarks, in AWS Glue. Job run history is accessible for 90 days for your workflow and job run.

Data types

- JobRun structure
- Predecessor structure
- JobBookmarkEntry structure
- BatchStopJobRunSuccessfulSubmission structure
- BatchStopJobRunError structure
- NotificationProperty structure

JobRun structure

Contains information about a job run.

Fields

- `Id` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of this job run.
- `Attempt` – Number (integer).

  The number of the attempt to run this job.
- `PreviousRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the previous run of this job. For example, the `JobRunId` specified in the `StartJobRun` action.
- `TriggerName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the trigger that started this job run.

- `JobName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the job definition being used in this run.

- `JobMode` – UTF-8 string (valid values: `SCRIPT=""` | `VISUAL=""` | `NOTEBOOK=""`).

  A mode that describes how a job was created. Valid values are:

  - `SCRIPT` - The job was created using the AWS Glue Studio script editor.
  - `VISUAL` - The job was created using the AWS Glue Studio visual editor.
  - `NOTEBOOK` - The job was created using an interactive sessions notebook.

  When the `JobMode` field is missing or null, `SCRIPT` is assigned as the default value.

- `JobRunQueuingEnabled` – Boolean.

  Specifies whether job run queuing is enabled for the job run.

  A value of true means job run queuing is enabled for the job run. If false or not populated, the job run will not be considered for queueing.

- `StartedOn` – Timestamp.

  The date and time at which this job run was started.

- `LastModifiedOn` – Timestamp.

  The last time that this job run was modified.

- `CompletedOn` – Timestamp.

  The date and time that this job run completed.

- `JobRunState` – UTF-8 string (valid values: `STARTING` | `RUNNING` | `STOPPING` | `STOPPED` | `SUCCEEDED` | `FAILED` | `TIMEOUT` | `ERROR` | `WAITING` | `EXPIRED`).

  The current state of the job run. For more information about the statuses of jobs that have terminated abnormally, see AWS Glue Job Run Statuses.

- `Arguments` – A map array of key-value pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

The job arguments associated with this run. For this job run, they replace the default arguments set in the job definition itself.

You can specify arguments here that your own job-execution script consumes, as well as arguments that AWS Glue itself consumes.

Job arguments may be logged. Do not pass plaintext secrets as arguments. Retrieve secrets from a AWS Glue Connection, AWS Secrets Manager or other secret management mechanism if you intend to keep them within the Job.

For information about how to specify and consume your own Job arguments, see the Calling AWS Glue APIs in Python topic in the developer guide.

For information about the arguments you can provide to this field when configuring Spark jobs, see the Special Parameters Used by AWS Glue topic in the developer guide.

For information about the arguments you can provide to this field when configuring Ray jobs, see Using job parameters in Ray jobs in the developer guide.

- `ErrorMessage` – UTF-8 string.

  An error message associated with this job run.
- `PredecessorRuns` – An array of Predecessor objects.

  A list of predecessors to this job run.
- `AllocatedCapacity` – Number (integer).

  This field is deprecated. Use `MaxCapacity` instead.

  The number of AWS Glue data processing units (DPUs) allocated to this JobRun. From 2 to 100 DPUs can be allocated; the default is 10. A DPU is a relative measure of processing power that consists of 4 vCPUs of compute capacity and 16 GB of memory. For more information, see the AWS Glue pricing page.
- `ExecutionTime` – Number (integer).

  The amount of time (in seconds) that the job run consumed resources.
- `Timeout` – Number (integer), at least 1.

The `JobRun` timeout in minutes. This is the maximum time that a job run can consume resources before it is terminated and enters `TIMEOUT` status. This value overrides the timeout value set in the parent job.

Jobs must have timeout values less than 7 days or 10080 minutes. Otherwise, the jobs will throw an exception.

When the value is left blank, the timeout is defaulted to 2880 minutes.

Any existing AWS Glue jobs that had a timeout value greater than 7 days will be defaulted to 7 days. For instance if you have specified a timeout of 20 days for a batch job, it will be stopped on the 7th day.

For streaming jobs, if you have set up a maintenance window, it will be restarted during the maintenance window after 7 days.

- `MaxCapacity` – Number (double).

  For Glue version 1.0 or earlier jobs, using the standard worker type, the number of AWS Glue data processing units (DPUs) that can be allocated when this job runs. A DPU is a relative measure of processing power that consists of 4 vCPUs of compute capacity and 16 GB of memory. For more information, see the AWS Glue pricing page.

  For Glue version 2.0+ jobs, you cannot specify a `Maximum capacity`. Instead, you should specify a `Worker type` and the `Number of workers`.

  Do not set `MaxCapacity` if using `WorkerType` and `NumberOfWorkers`.

  The value that can be allocated for `MaxCapacity` depends on whether you are running a Python shell job, an Apache Spark ETL job, or an Apache Spark streaming ETL job:

  - When you specify a Python shell job (`JobCommand.Name`="pythonshell"), you can allocate either 0.0625 or 1 DPU. The default is 0.0625 DPU.
  - When you specify an Apache Spark ETL job (`JobCommand.Name`="glueetl") or Apache Spark streaming ETL job (`JobCommand.Name`="gluestreaming"), you can allocate from 2 to 100 DPUs. The default is 10 DPUs. This job type cannot have a fractional DPU allocation.
- `WorkerType` – UTF-8 string (valid values: `Standard=""` | `G.1X=""` | `G.2X=""` | `G.025X=""` | `G.4X=""` | `G.8X=""` | `Z.2X=""`).

The type of predefined worker that is allocated when a job runs. Accepts a value of G.1X, G.2X, G.4X, G.8X or G.025X for Spark jobs. Accepts the value Z.2X for Ray jobs.

- For the `G.1X` worker type, each worker maps to 1 DPU (4 vCPUs, 16 GB of memory) with 94GB disk, and provides 1 executor per worker. We recommend this worker type for workloads such as data transforms, joins, and queries, to offers a scalable and cost effective way to run most jobs.
- For the `G.2X` worker type, each worker maps to 2 DPU (8 vCPUs, 32 GB of memory) with 138GB disk, and provides 1 executor per worker. We recommend this worker type for workloads such as data transforms, joins, and queries, to offers a scalable and cost effective way to run most jobs.
- For the `G.4X` worker type, each worker maps to 4 DPU (16 vCPUs, 64 GB of memory) with 256GB disk, and provides 1 executor per worker. We recommend this worker type for jobs whose workloads contain your most demanding transforms, aggregations, joins, and queries. This worker type is available only for AWS Glue version 3.0 or later Spark ETL jobs in the following AWS Regions: US East (Ohio), US East (N. Virginia), US West (Oregon), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Canada (Central), Europe (Frankfurt), Europe (Ireland), and Europe (Stockholm).
- For the `G.8X` worker type, each worker maps to 8 DPU (32 vCPUs, 128 GB of memory) with 512GB disk, and provides 1 executor per worker. We recommend this worker type for jobs whose workloads contain your most demanding transforms, aggregations, joins, and queries. This worker type is available only for AWS Glue version 3.0 or later Spark ETL jobs, in the same AWS Regions as supported for the `G.4X` worker type.
- For the `G.025X` worker type, each worker maps to 0.25 DPU (2 vCPUs, 4 GB of memory) with 84GB disk, and provides 1 executor per worker. We recommend this worker type for low volume streaming jobs. This worker type is only available for AWS Glue version 3.0 or later streaming jobs.
- For the `Z.2X` worker type, each worker maps to 2 M-DPU (8vCPUs, 64 GB of memory) with 128 GB disk, and provides up to 8 Ray workers based on the autoscaler.

- `NumberOfWorkers` – Number (integer).

  The number of workers of a defined `workerType` that are allocated when a job runs.

- `SecurityConfiguration` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the `SecurityConfiguration` structure to be used with this job run.

- `LogGroupName` – UTF-8 string.

  The name of the log group for secure logging that can be server-side encrypted in Amazon CloudWatch using AWS KMS. This name can be `/aws-glue/jobs/`, in which case the default encryption is `NONE`. If you add a role name and `SecurityConfiguration` name (in other words, `/aws-glue/jobs-yourRoleName-yourSecurityConfigurationName/`), then that security configuration is used to encrypt the log group.
- `NotificationProperty` – A NotificationProperty object.

  Specifies configuration properties of a job run notification.
- `GlueVersion` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #47.

  In Spark jobs, `GlueVersion` determines the versions of Apache Spark and Python that AWS Glue available in a job. The Python version indicates the version supported for jobs of type Spark.

  Ray jobs should set `GlueVersion` to `4.0` or greater. However, the versions of Ray, Python and additional libraries available in your Ray job are determined by the `Runtime` parameter of the Job command.

  For more information about the available AWS Glue versions and corresponding Spark and Python versions, see Glue version in the developer guide.

  Jobs that are created without specifying a Glue version default to Glue 0.9.
- `DPUSeconds` – Number (double).

  This field can be set for either job runs with execution class `FLEX` or when Auto Scaling is enabled, and represents the total time each executor ran during the lifecycle of a job run in seconds, multiplied by a DPU factor (1 for `G.1X`, 2 for `G.2X`, or 0.25 for `G.025X` workers). This value may be different than the `executionEngineRuntime` * `MaxCapacity` as in the case of Auto Scaling jobs, as the number of executors running at a given time may be less than the `MaxCapacity`. Therefore, it is possible that the value of `DPUSeconds` is less than `executionEngineRuntime` * `MaxCapacity`.
- `ExecutionClass` – UTF-8 string, not more than 16 bytes long (valid values: `FLEX=""` | `STANDARD=""`).

Indicates whether the job is run with a standard or flexible execution class. The standard execution-class is ideal for time-sensitive workloads that require fast job startup and dedicated resources.

The flexible execution class is appropriate for time-insensitive jobs whose start and completion times may vary.

Only jobs with AWS Glue version 3.0 and above and command type `glueetl` will be allowed to set `ExecutionClass` to `FLEX`. The flexible execution class is available for Spark jobs.

- `MaintenanceWindow` – UTF-8 string, matching the Custom string pattern #34.

  This field specifies a day of the week and hour for a maintenance window for streaming jobs. AWS Glue periodically performs maintenance activities. During these maintenance windows, AWS Glue will need to restart your streaming jobs.

  AWS Glue will restart the job within 3 hours of the specified maintenance window. For instance, if you set up the maintenance window for Monday at 10:00AM GMT, your jobs will be restarted between 10:00AM GMT to 1:00PM GMT.
- `ProfileName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of an AWS Glue usage profile associated with the job run.
- `StateDetail` – UTF-8 string, not more than 400000 bytes long.

  This field holds details that pertain to the state of a job run. The field is nullable.

  For example, when a job run is in a WAITING state as a result of job run queuing, the field has the reason why the job run is in that state.
- `ExecutionRoleSessionPolicy` – UTF-8 string, not less than 2 or more than 2048 bytes long.

  This inline session policy to the StartJobRun API allows you to dynamically restrict the permissions of the specified execution role for the scope of the job, without requiring the creation of additional IAM roles.

Predecessor structure

A job run that was used in the predicate of a conditional trigger that triggered this job run.

Fields

- `JobName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the job definition used by the predecessor job run.
- `RunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The job-run ID of the predecessor job run.

JobBookmarkEntry structure

Defines a point that a job can resume processing.

Fields

- `JobName` – UTF-8 string.

  The name of the job in question.
- `Version` – Number (integer).

  The version of the job.
- `Run` – Number (integer).

  The run ID number.
- `Attempt` – Number (integer).

  The attempt ID number.
- `PreviousRunId` – UTF-8 string.

  The unique run identifier associated with the previous job run.
- `RunId` – UTF-8 string.

  The run ID number.
- `JobBookmark` – UTF-8 string.

  The bookmark itself.

BatchStopJobRunSuccessfulSubmission structure

Records a successful request to stop a specified `JobRun`.

Fields

- `JobName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the job definition used in the job run that was stopped.
- `JobRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The `JobRunId` of the job run that was stopped.

BatchStopJobRunError structure

Records an error that occurred when attempting to stop a specified job run.

Fields

- `JobName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the job definition that is used in the job run in question.
- `JobRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The `JobRunId` of the job run in question.
- `ErrorDetail` – An ErrorDetail object.

  Specifies details about the error that was encountered.

NotificationProperty structure

Specifies configuration properties of a notification.

Fields

- `NotifyDelayAfter` – Number (integer), at least 1.

After a job run starts, the number of minutes to wait before sending a job run delay notification.

Operations

- StartJobRun action (Python: start_job_run)
- BatchStopJobRun action (Python: batch_stop_job_run)
- GetJobRun action (Python: get_job_run)
- GetJobRuns action (Python: get_job_runs)
- GetJobBookmark action (Python: get_job_bookmark)
- GetJobBookmarks action (Python: get_job_bookmarks)
- ResetJobBookmark action (Python: reset_job_bookmark)

StartJobRun action (Python: start_job_run)

Starts a job run using a job definition.

Request

- `JobName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the job definition to use.
- `JobRunQueuingEnabled` – Boolean.

  Specifies whether job run queuing is enabled for the job run.

  A value of true means job run queuing is enabled for the job run. If false or not populated, the job run will not be considered for queueing.
- `JobRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of a previous `JobRun` to retry.
- `Arguments` – A map array of key-value pairs.

  Each key is a UTF-8 string.

Each value is a UTF-8 string.

The job arguments associated with this run. For this job run, they replace the default arguments set in the job definition itself.

You can specify arguments here that your own job-execution script consumes, as well as arguments that AWS Glue itself consumes.

Job arguments may be logged. Do not pass plaintext secrets as arguments. Retrieve secrets from a AWS Glue Connection, AWS Secrets Manager or other secret management mechanism if you intend to keep them within the Job.

For information about how to specify and consume your own Job arguments, see the Calling AWS Glue APIs in Python topic in the developer guide.

For information about the arguments you can provide to this field when configuring Spark jobs, see the Special Parameters Used by AWS Glue topic in the developer guide.

For information about the arguments you can provide to this field when configuring Ray jobs, see Using job parameters in Ray jobs in the developer guide.

- `AllocatedCapacity` – Number (integer).

  This field is deprecated. Use `MaxCapacity` instead.

  The number of AWS Glue data processing units (DPUs) to allocate to this JobRun. You can allocate a minimum of 2 DPUs; the default is 10. A DPU is a relative measure of processing power that consists of 4 vCPUs of compute capacity and 16 GB of memory. For more information, see the AWS Glue pricing page.
- `Timeout` – Number (integer), at least 1.

  The `JobRun` timeout in minutes. This is the maximum time that a job run can consume resources before it is terminated and enters `TIMEOUT` status. This value overrides the timeout value set in the parent job.

  Jobs must have timeout values less than 7 days or 10080 minutes. Otherwise, the jobs will throw an exception.

  When the value is left blank, the timeout is defaulted to 2880 minutes.

Any existing AWS Glue jobs that had a timeout value greater than 7 days will be defaulted to 7 days. For instance if you have specified a timeout of 20 days for a batch job, it will be stopped on the 7th day.

For streaming jobs, if you have set up a maintenance window, it will be restarted during the maintenance window after 7 days.

- `MaxCapacity` – Number (double).

  For Glue version 1.0 or earlier jobs, using the standard worker type, the number of AWS Glue data processing units (DPUs) that can be allocated when this job runs. A DPU is a relative measure of processing power that consists of 4 vCPUs of compute capacity and 16 GB of memory. For more information, see the AWS Glue pricing page.

  For Glue version 2.0+ jobs, you cannot specify a `Maximum capacity`. Instead, you should specify a `Worker type` and the `Number of workers`.

  Do not set `MaxCapacity` if using `WorkerType` and `NumberOfWorkers`.

  The value that can be allocated for `MaxCapacity` depends on whether you are running a Python shell job, an Apache Spark ETL job, or an Apache Spark streaming ETL job:

  - When you specify a Python shell job (`JobCommand.Name`="pythonshell"), you can allocate either 0.0625 or 1 DPU. The default is 0.0625 DPU.
  - When you specify an Apache Spark ETL job (`JobCommand.Name`="glueetl") or Apache Spark streaming ETL job (`JobCommand.Name`="gluestreaming"), you can allocate from 2 to 100 DPUs. The default is 10 DPUs. This job type cannot have a fractional DPU allocation.

- `SecurityConfiguration` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the `SecurityConfiguration` structure to be used with this job run.

- `NotificationProperty` – A NotificationProperty object.

  Specifies configuration properties of a job run notification.

- `WorkerType` – UTF-8 string (valid values: `Standard=""` | `G.1X=""` | `G.2X=""` | `G.025X=""` | `G.4X=""` | `G.8X=""` | `Z.2X=""`).

  The type of predefined worker that is allocated when a job runs. Accepts a value of G.1X, G.2X, G.4X, G.8X or G.025X for Spark jobs. Accepts the value Z.2X for Ray jobs.

  - For the `G.1X` worker type, each worker maps to 1 DPU (4 vCPUs, 16 GB of memory) with 94GB disk, and provides 1 executor per worker. We recommend this worker type for workloads such as data transforms, joins, and queries, to offers a scalable and cost effective way to run most jobs.
  - For the `G.2X` worker type, each worker maps to 2 DPU (8 vCPUs, 32 GB of memory) with 138GB disk, and provides 1 executor per worker. We recommend this worker type for workloads such as data transforms, joins, and queries, to offers a scalable and cost effective way to run most jobs.
  - For the `G.4X` worker type, each worker maps to 4 DPU (16 vCPUs, 64 GB of memory) with 256GB disk, and provides 1 executor per worker. We recommend this worker type for jobs whose workloads contain your most demanding transforms, aggregations, joins, and queries. This worker type is available only for AWS Glue version 3.0 or later Spark ETL jobs in the following AWS Regions: US East (Ohio), US East (N. Virginia), US West (Oregon), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Canada (Central), Europe (Frankfurt), Europe (Ireland), and Europe (Stockholm).
  - For the `G.8X` worker type, each worker maps to 8 DPU (32 vCPUs, 128 GB of memory) with 512GB disk, and provides 1 executor per worker. We recommend this worker type for jobs whose workloads contain your most demanding transforms, aggregations, joins, and queries. This worker type is available only for AWS Glue version 3.0 or later Spark ETL jobs, in the same AWS Regions as supported for the `G.4X` worker type.
  - For the `G.025X` worker type, each worker maps to 0.25 DPU (2 vCPUs, 4 GB of memory) with 84GB disk, and provides 1 executor per worker. We recommend this worker type for low volume streaming jobs. This worker type is only available for AWS Glue version 3.0 or later streaming jobs.
  - For the `Z.2X` worker type, each worker maps to 2 M-DPU (8vCPUs, 64 GB of memory) with 128 GB disk, and provides up to 8 Ray workers based on the autoscaler.
- `NumberOfWorkers` – Number (integer).

  The number of workers of a defined `workerType` that are allocated when a job runs.
- `ExecutionClass` – UTF-8 string, not more than 16 bytes long (valid values: `FLEX=""` | `STANDARD=""`).

  Indicates whether the job is run with a standard or flexible execution class. The standard execution-class is ideal for time-sensitive workloads that require fast job startup and dedicated resources.

The flexible execution class is appropriate for time-insensitive jobs whose start and completion times may vary.

Only jobs with AWS Glue version 3.0 and above and command type `glueetl` will be allowed to set `ExecutionClass` to `FLEX`. The flexible execution class is available for Spark jobs.

- `ProfileName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of an AWS Glue usage profile associated with the job run.
- `ExecutionRoleSessionPolicy` – UTF-8 string, not less than 2 or more than 2048 bytes long.

  This inline session policy to the StartJobRun API allows you to dynamically restrict the permissions of the specified execution role for the scope of the job, without requiring the creation of additional IAM roles.

Response

- `JobRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID assigned to this job run.

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`
- `ResourceNumberLimitExceededException`
- `ConcurrentRunsExceededException`

BatchStopJobRun action (Python: batch_stop_job_run)

Stops one or more job runs for a specified job definition.

Request

- `JobName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the job definition for which to stop job runs.
- `JobRunIds` – *Required:* An array of UTF-8 strings, not less than 1 or more than 25 strings.

  A list of the `JobRunIds` that should be stopped for that job definition.

Response

- `SuccessfulSubmissions` – An array of BatchStopJobRunSuccessfulSubmission objects.

  A list of the JobRuns that were successfully submitted for stopping.
- `Errors` – An array of BatchStopJobRunError objects.

  A list of the errors that were encountered in trying to stop `JobRuns`, including the `JobRunId` for which each error was encountered and details about the error.

Errors

- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`

GetJobRun action (Python: get_job_run)

Retrieves the metadata for a given job run. Job run history is accessible for 365 days for your workflow and job run.

Request

- `JobName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Name of the job definition being run.

- `RunId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the job run.
- `PredecessorsIncluded` – Boolean.

  True if a list of predecessor runs should be returned.

Response

- `JobRun` – A JobRun object.

  The requested job-run metadata.

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`

GetJobRuns action (Python: get_job_runs)

Retrieves metadata for all runs of a given job definition.

`GetJobRuns` returns the job runs in chronological order, with the newest jobs returned first.

Request

- `JobName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the job definition for which to retrieve all job runs.
- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation call.
- `MaxResults` – Number (integer), not less than 1 or more than 200.

The maximum size of the response.

Response

- `JobRuns` – An array of JobRun objects.

  A list of job-run metadata objects.
- `NextToken` – UTF-8 string.

  A continuation token, if not all requested job runs have been returned.

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`

GetJobBookmark action (Python: get_job_bookmark)

Returns information on a job bookmark entry.

For more information about enabling and using job bookmarks, see:

- Tracking processed data using job bookmarks
- Job parameters used by AWS Glue
- Job structure

Request

- `JobName` – *Required:* UTF-8 string.

  The name of the job in question.
- `Version` – Number (integer).

  The version of the job.

- `RunId` – UTF-8 string.

  The unique run identifier associated with this job run.

Response

- `JobBookmarkEntry` – A JobBookmarkEntry object.

  A structure that defines a point that a job can resume processing.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `ValidationException`

GetJobBookmarks action (Python: get_job_bookmarks)

Returns information on the job bookmark entries. The list is ordered on decreasing version numbers.

For more information about enabling and using job bookmarks, see:

- Tracking processed data using job bookmarks
- Job parameters used by AWS Glue
- Job structure

Request

- `JobName` – *Required:* UTF-8 string.

  The name of the job in question.
- `MaxResults` – Number (integer).

  The maximum size of the response.

- `NextToken` – Number (integer).

  A continuation token, if this is a continuation call.

Response

- `JobBookmarkEntries` – An array of JobBookmarkEntry objects.

  A list of job bookmark entries that defines a point that a job can resume processing.
- `NextToken` – Number (integer).

  A continuation token, which has a value of 1 if all the entries are returned, or > 1 if not all requested job runs have been returned.

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`

ResetJobBookmark action (Python: reset_job_bookmark)

Resets a bookmark entry.

For more information about enabling and using job bookmarks, see:

- Tracking processed data using job bookmarks
- Job parameters used by AWS Glue
- Job structure

Request

- `JobName` – *Required:* UTF-8 string.

  The name of the job in question.
- `RunId` – UTF-8 string.

The unique run identifier associated with this job run.

Response

- `JobBookmarkEntry` – A JobBookmarkEntry object.

  The reset bookmark entry.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`

Triggers

The Triggers API describes the data types and API related to creating, updating, or deleting, and starting and stopping job triggers in AWS Glue.

Data types

- Trigger structure
- TriggerUpdate structure
- Predicate structure
- Condition structure
- Action structure
- EventBatchingCondition structure

Trigger structure

Information about a specific trigger.

Fields

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the trigger.
- `WorkflowName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the workflow associated with the trigger.
- `Id` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Reserved for future use.
- `Type` – UTF-8 string (valid values: `SCHEDULED` | `CONDITIONAL` | `ON_DEMAND` | `EVENT`).

  The type of trigger that this is.
- `State` – UTF-8 string (valid values: `CREATING` | `CREATED` | `ACTIVATING` | `ACTIVATED` | `DEACTIVATING` | `DEACTIVATED` | `DELETING` | `UPDATING`).

  The current state of the trigger.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of this trigger.
- `Schedule` – UTF-8 string.

  A `cron` expression used to specify the schedule (see Time-Based Schedules for Jobs and Crawlers. For example, to run something every day at 12:15 UTC, you would specify: `cron(15 12 * * ? *)`.
- `Actions` – An array of Action objects.

  The actions initiated by this trigger.
- `Predicate` – A Predicate object.

  The predicate of this trigger, which defines when it will fire.
- `EventBatchingCondition` – An EventBatchingCondition object.

Batch condition that must be met (specified number of events received or batch time window expired) before EventBridge event trigger fires.

TriggerUpdate structure

A structure used to provide information used to update a trigger. This object updates the previous trigger definition by overwriting it completely.

Fields

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Reserved for future use.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of this trigger.
- `Schedule` – UTF-8 string.

  A `cron` expression used to specify the schedule (see Time-Based Schedules for Jobs and Crawlers. For example, to run something every day at 12:15 UTC, you would specify: `cron(15 12 * * ? *)`.
- `Actions` – An array of Action objects.

  The actions initiated by this trigger.
- `Predicate` – A Predicate object.

  The predicate of this trigger, which defines when it will fire.
- `EventBatchingCondition` – An EventBatchingCondition object.

  Batch condition that must be met (specified number of events received or batch time window expired) before EventBridge event trigger fires.

Predicate structure

Defines the predicate of the trigger, which determines when it fires.

Fields

- `Logical` – UTF-8 string (valid values: `AND` | `ANY`).

  An optional field if only one condition is listed. If multiple conditions are listed, then this field is required.
- `Conditions` – An array of Condition objects, not more than 500 structures.

  A list of the conditions that determine when the trigger will fire.

Condition structure

Defines a condition under which a trigger fires.

Fields

- `LogicalOperator` – UTF-8 string (valid values: `EQUALS`).

  A logical operator.
- `JobName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the job whose `JobRuns` this condition applies to, and on which this trigger waits.
- `State` – UTF-8 string (valid values: `STARTING` | `RUNNING` | `STOPPING` | `STOPPED` | `SUCCEEDED` | `FAILED` | `TIMEOUT` | `ERROR` | `WAITING` | `EXPIRED`).

  The condition state. Currently, the only job states that a trigger can listen for are `SUCCEEDED`, `STOPPED`, `FAILED`, and `TIMEOUT`. The only crawler states that a trigger can listen for are `SUCCEEDED`, `FAILED`, and `CANCELLED`.
- `CrawlerName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the crawler to which this condition applies.
- `CrawlState` – UTF-8 string (valid values: `RUNNING` | `CANCELLING` | `CANCELLED` | `SUCCEEDED` | `FAILED` | `ERROR`).

  The state of the crawler to which this condition applies.

Action structure

Defines an action to be initiated by a trigger.

Fields

- `JobName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of a job to be run.
- `Arguments` – A map array of key-value pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

  The job arguments used when this trigger fires. For this job run, they replace the default arguments set in the job definition itself.

  You can specify arguments here that your own job-execution script consumes, as well as arguments that AWS Glue itself consumes.

  For information about how to specify and consume your own Job arguments, see the Calling AWS Glue APIs in Python topic in the developer guide.

  For information about the key-value pairs that AWS Glue consumes to set up your job, see the Special Parameters Used by AWS Glue topic in the developer guide.
- `Timeout` – Number (integer), at least 1.

  The `JobRun` timeout in minutes. This is the maximum time that a job run can consume resources before it is terminated and enters `TIMEOUT` status. This overrides the timeout value set in the parent job.

  Jobs must have timeout values less than 7 days or 10080 minutes. Otherwise, the jobs will throw an exception.

  When the value is left blank, the timeout is defaulted to 2880 minutes.

  Any existing AWS Glue jobs that had a timeout value greater than 7 days will be defaulted to 7 days. For instance if you have specified a timeout of 20 days for a batch job, it will be stopped on the 7th day.

For streaming jobs, if you have set up a maintenance window, it will be restarted during the maintenance window after 7 days.

- `SecurityConfiguration` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the `SecurityConfiguration` structure to be used with this action.
- `NotificationProperty` – A NotificationProperty object.

  Specifies configuration properties of a job run notification.
- `CrawlerName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the crawler to be used with this action.

EventBatchingCondition structure

Batch condition that must be met (specified number of events received or batch time window expired) before EventBridge event trigger fires.

Fields

- `BatchSize` – *Required:* Number (integer), not less than 1 or more than 100.

  Number of events that must be received from Amazon EventBridge before EventBridge event trigger fires.
- `BatchWindow` – Number (integer), not less than 1 or more than 900.

  Window of time in seconds after which EventBridge event trigger fires. Window starts when first event is received.

Operations

- CreateTrigger action (Python: create_trigger)
- StartTrigger action (Python: start_trigger)
- GetTrigger action (Python: get_trigger)
- GetTriggers action (Python: get_triggers)

- UpdateTrigger action (Python: update_trigger)
- StopTrigger action (Python: stop_trigger)
- DeleteTrigger action (Python: delete_trigger)
- ListTriggers action (Python: list_triggers)
- BatchGetTriggers action (Python: batch_get_triggers)

CreateTrigger action (Python: create_trigger)

Creates a new trigger.

Job arguments may be logged. Do not pass plaintext secrets as arguments. Retrieve secrets from a AWS Glue Connection, AWS Secrets Manager or other secret management mechanism if you intend to keep them within the Job.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the trigger.
- `WorkflowName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the workflow associated with the trigger.
- `Type` – *Required:* UTF-8 string (valid values: `SCHEDULED` | `CONDITIONAL` | `ON_DEMAND` | `EVENT`).

  The type of the new trigger.
- `Schedule` – UTF-8 string.

  A `cron` expression used to specify the schedule (see Time-Based Schedules for Jobs and Crawlers. For example, to run something every day at 12:15 UTC, you would specify: `cron(15 12 * * ? *)`.

  This field is required when the trigger type is SCHEDULED.
- `Predicate` – A Predicate object.

  A predicate to specify when the new trigger should fire.

This field is required when the trigger type is `CONDITIONAL`.

- `Actions` – *Required:* An array of [Action](#) objects.

  The actions initiated by this trigger when it fires.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the new trigger.
- `StartOnCreation` – Boolean.

  Set to `true` to start `SCHEDULED` and `CONDITIONAL` triggers when created. True is not supported for `ON_DEMAND` triggers.
- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  The tags to use with this trigger. You may use tags to limit access to the trigger. For more information about tags in AWS Glue, see AWS Tags in AWS Glue in the developer guide.
- `EventBatchingCondition` – An EventBatchingCondition object.

  Batch condition that must be met (specified number of events received or batch time window expired) before EventBridge event trigger fires.

Response

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the trigger.

Errors

- `AlreadyExistsException`
- `EntityNotFoundException`
- `InvalidInputException`

- `IdempotentParameterMismatchException`
- `InternalServiceException`
- `OperationTimeoutException`
- `ResourceNumberLimitExceededException`
- `ConcurrentModificationException`

StartTrigger action (Python: start_trigger)

Starts an existing trigger. See Triggering Jobs for information about how different types of trigger are started.

Request

- Name – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the trigger to start.

Response

- Name – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the trigger that was started.

Errors

- `InvalidInputException`
- `InternalServiceException`
- `EntityNotFoundException`
- `OperationTimeoutException`
- `ResourceNumberLimitExceededException`
- `ConcurrentRunsExceededException`

GetTrigger action (Python: get_trigger)

Retrieves the definition of a trigger.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the trigger to retrieve.

Response

- `Trigger` – A Trigger object.

  The requested trigger definition.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`

GetTriggers action (Python: get_triggers)

Gets all the triggers associated with a job.

Request

- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation call.
- `DependentJobName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the job to retrieve triggers for. The trigger that can start this job is returned, and if there is no such trigger, all triggers are returned.

- `MaxResults` – Number (integer), not less than 1 or more than 200.

  The maximum size of the response.

Response

- `Triggers` – An array of [Trigger](#) objects.

  A list of triggers for the specified job.
- `NextToken` – UTF-8 string.

  A continuation token, if not all the requested triggers have yet been returned.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`

UpdateTrigger action (Python: update_trigger)

Updates a trigger definition.

Job arguments may be logged. Do not pass plaintext secrets as arguments. Retrieve secrets from a AWS Glue Connection, AWS Secrets Manager or other secret management mechanism if you intend to keep them within the Job.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the [Single-line string pattern](#).

  The name of the trigger to update.
- `TriggerUpdate` – *Required:* A [TriggerUpdate](#) object.

  The new values with which to update the trigger.

Response

- `Trigger` – A Trigger object.

  The resulting trigger definition.

Errors

- `InvalidInputException`
- `InternalServiceException`
- `EntityNotFoundException`
- `OperationTimeoutException`
- `ConcurrentModificationException`

StopTrigger action (Python: stop_trigger)

Stops a specified trigger.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the trigger to stop.

Response

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the trigger that was stopped.

Errors

- `InvalidInputException`
- `InternalServiceException`
- `EntityNotFoundException`

- `OperationTimeoutException`
- `ConcurrentModificationException`

DeleteTrigger action (Python: delete_trigger)

Deletes a specified trigger. If the trigger is not found, no exception is thrown.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the trigger to delete.

Response

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the trigger that was deleted.

Errors

- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `ConcurrentModificationException`

ListTriggers action (Python: list_triggers)

Retrieves the names of all trigger resources in this AWS account, or the resources with the specified tag. This operation allows you to see which resources are available in your account, and their names.

This operation takes the optional `Tags` field, which you can use as a filter on the response so that tagged resources can be retrieved as a group. If you choose to use tags filtering, only resources with the tag are retrieved.

Request

- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation request.
- `DependentJobName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the job for which to retrieve triggers. The trigger that can start this job is returned. If there is no such trigger, all triggers are returned.
- `MaxResults` – Number (integer), not less than 1 or more than 200.

  The maximum size of a list to return.
- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  Specifies to return only these tagged resources.

Response

- `TriggerNames` – An array of UTF-8 strings.

  The names of all triggers in the account, or the triggers with the specified tags.
- `NextToken` – UTF-8 string.

  A continuation token, if the returned list does not contain the last metric available.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`

BatchGetTriggers action (Python: batch_get_triggers)

Returns a list of resource metadata for a given list of trigger names. After calling the `ListTriggers` operation, you can call this operation to access the data to which you have been granted permissions. This operation supports all IAM permissions, including permission conditions that uses tags.

Request

- `TriggerNames` – *Required:* An array of UTF-8 strings.

  A list of trigger names, which may be the names returned from the `ListTriggers` operation.

Response

- `Triggers` – An array of Trigger objects.

  A list of trigger definitions.
- `TriggersNotFound` – An array of UTF-8 strings.

  A list of names of triggers not found.

Errors

- `InternalServiceException`
- `OperationTimeoutException`
- `InvalidInputException`

Integration APIs in AWS Glue

Data types

- Integration structure
- IntegrationConfig structure
- IntegrationPartition structure
- IntegrationError structure

- IntegrationFilter structure
- InboundIntegration structure
- SourceProcessingProperties structure
- TargetProcessingProperties structure
- SourceTableConfig structure
- TargetTableConfig structure

Integration structure

Describes a zero-ETL integration.

Fields

- `SourceArn` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The ARN for the source of the integration.
- `TargetArn` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The ARN for the target of the integration.
- `Description` – UTF-8 string, not more than 1000 bytes long, matching the Custom string pattern #12.

  A description for the integration.
- `IntegrationName` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  A unique name for the integration.
- `IntegrationArn` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The Amazon Resource Name (ARN) for the integration.
- `KmsKeyId` – UTF-8 string, not less than 1 or more than 2048 bytes long.

  The ARN of a KMS key used for encrypting the channel.
- `AdditionalEncryptionContext` – A map array of key-value pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

An optional set of non-secret key–value pairs that contains additional contextual information for encryption. This can only be provided if `KMSKeyId` is provided.

- `Tags` – An array of [Tag](#) objects.

  Metadata assigned to the resource consisting of a list of key-value pairs.
- `Status` – *Required:* UTF-8 string (valid values: `CREATING` | `ACTIVE` | `MODIFYING` | `FAILED` | `DELETING` | `SYNCING` | `NEEDS_ATTENTION`).

  The possible statuses are:
  - CREATING: The integration is being created.
  - ACTIVE: The integration creation succeeds.
  - MODIFYING: The integration is being modified.
  - FAILED: The integration creation fails.
  - DELETING: The integration is deleted.
  - SYNCING: The integration is synchronizing.
  - NEEDS_ATTENTION: The integration needs attention, such as synchronization.
- `CreateTime` – *Required:* Timestamp.

  The time that the integration was created, in UTC.
- `IntegrationConfig` – An [IntegrationConfig](#) object.

  Properties associated with the integration.
- `Errors` – An array of [IntegrationError](#) objects.

  A list of errors associated with the integration.
- `DataFilter` – UTF-8 string, not less than 1 or more than 2048 bytes long.

  Selects source tables for the integration using Maxwell filter syntax.

IntegrationConfig structure

Properties associated with the integration.

Fields

- `RefreshInterval` – UTF-8 string, not less than 1 or more than 128 bytes long.

Specifies the frequency at which CDC (Change Data Capture) pulls or incremental loads should occur. This parameter provides flexibility to align the refresh rate with your specific data update patterns, system load considerations, and performance optimization goals. Time increment can be set from 15 minutes to 8640 minutes (six days).

- `SourceProperties` – A map array of key-value pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

  A collection of key-value pairs that specify additional properties for the integration source. These properties provide configuration options that can be used to customize the behavior of the ODB source during data integration operations.
- `ContinuousSync` – Boolean.

  Enables continuous synchronization for on-demand data extractions from SaaS applications to AWS data services like Amazon Redshift and Amazon S3.

IntegrationPartition structure

A structure that describes how data is partitioned on the target.

Fields

- `FieldName` – UTF-8 string, not less than 1 or more than 128 bytes long.

  The field name used to partition data on the target. Avoid using columns that have unique values for each row (for example, 'LastModifiedTimestamp', 'SystemModTimeStamp') as the partition column. These columns are not suitable for partitioning because they create a large number of small partitions, which can lead to performance issues.
- `FunctionSpec` – UTF-8 string, not less than 1 or more than 128 bytes long.

  Specifies the function used to partition data on the target. The accepted values for this parameter are:
  - `identity` - Uses source values directly without transformation
  - `year` - Extracts the year from timestamp values (e.g., 2023)
  - `month` - Extracts the month from timestamp values (e.g., 2023-01)

  - `day` - Extracts the day from timestamp values (e.g., 2023-01-15)
  - `hour` - Extracts the hour from timestamp values (e.g., 2023-01-15-14)
- `ConversionSpec` – UTF-8 string, not less than 1 or more than 128 bytes long.

  Specifies the timestamp format of the source data. Valid values are:

  - `epoch_sec` - Unix epoch timestamp in seconds
  - `epoch_milli` - Unix epoch timestamp in milliseconds
  - `iso` - ISO 8601 formatted timestamp

> **Note**
>
> Only specify `ConversionSpec` when using timestamp-based partition functions (year, month, day, or hour). AWS Glue Zero-ETL uses this parameter to correctly transform source data into timestamp format before partitioning.
> Do not use high-cardinality columns with the `identity` partition function. High-cardinality columns include:
>
> - Primary keys
> - Timestamp fields (such as `LastModifiedTimestamp`, `CreatedDate`)
> - System-generated timestamps
>
> Using high-cardinality columns with identity partitioning creates many small partitions, which can significantly degrade ingestion performance.

IntegrationError structure

An error associated with a zero-ETL integration.

Fields

- `ErrorCode` – UTF-8 string, not less than 1 or more than 128 bytes long.

  The code associated with this error.

- `ErrorMessage` – UTF-8 string, not less than 1 or more than 2048 bytes long.

  A message describing the error.

IntegrationFilter structure

A filter that can be used when invoking a `DescribeIntegrations` request.

Fields

- `Name` – UTF-8 string, not less than 1 or more than 128 bytes long.

  The name of the filter.
- `Values` – An array of UTF-8 strings.

  A list of filter values.

InboundIntegration structure

A structure for an integration that writes data into a resource.

Fields

- `SourceArn` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The ARN of the source resource for the integration.
- `TargetArn` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The ARN of the target resource for the integration.
- `IntegrationArn` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The ARN of the zero-ETL integration.
- `Status` – *Required:* UTF-8 string (valid values: `CREATING` | `ACTIVE` | `MODIFYING` | `FAILED` | `DELETING` | `SYNCING` | `NEEDS_ATTENTION`).

  The possible statuses are:
  - CREATING: The integration is being created.
  - ACTIVE: The integration creation succeeds.
  - MODIFYING: The integration is being modified.
  - FAILED: The integration creation fails.
  - DELETING: The integration is deleted.
  - SYNCING: The integration is synchronizing.

  - NEEDS_ATTENTION: The integration needs attention, such as synchronization.
- `CreateTime` – *Required:* Timestamp.

  The time that the integration was created, in UTC.
- `IntegrationConfig` – An IntegrationConfig object.

  Properties associated with the integration.
- `Errors` – An array of IntegrationError objects.

  A list of errors associated with the integration.

SourceProcessingProperties structure

The resource properties associated with the integration source.

Fields

- `RoleArn` – UTF-8 string, not less than 1 or more than 128 bytes long.

  The IAM role to access the AWS Glue connection.

TargetProcessingProperties structure

The resource properties associated with the integration target.

Fields

- `RoleArn` – UTF-8 string, not less than 1 or more than 128 bytes long.

  The IAM role to access the AWS Glue database.
- `KmsArn` – UTF-8 string, not less than 1 or more than 2048 bytes long.

  The ARN of the KMS key used for encryption.
- `ConnectionName` – UTF-8 string, not less than 1 or more than 128 bytes long.

  The AWS Glue network connection to configure the AWS Glue job running in the customer VPC.
- `EventBusArn` – UTF-8 string, not less than 1 or more than 2048 bytes long.

  The ARN of an Eventbridge event bus to receive the integration status notification.

SourceTableConfig structure

Properties used by the source leg to process data from the source.

Fields

- `Fields` – An array of UTF-8 strings.

  A list of fields used for column-level filtering. Currently unsupported.
- `FilterPredicate` – UTF-8 string, not less than 1 or more than 128 bytes long.

  A condition clause used for row-level filtering. Currently unsupported.
- `PrimaryKey` – An array of UTF-8 strings.

  Provide the primary key set for this table. Currently supported specifically for SAP `EntityOf` entities upon request. Contact AWS Support to make this feature available.
- `RecordUpdateField` – UTF-8 string, not less than 1 or more than 128 bytes long.

  Incremental pull timestamp-based field. Currently unsupported.

TargetTableConfig structure

Properties used by the target leg to partition the data on the target.

Fields

- `UnnestSpec` – UTF-8 string (valid values: `TOPLEVEL` | `FULL` | `NOUNNEST`).

  Specifies how nested objects are flattened to top-level elements. Valid values are: "TOPLEVEL", "FULL", or "NOUNNEST".
- `PartitionSpec` – An array of IntegrationPartition objects.

  Determines the file layout on the target.
- `TargetTableName` – UTF-8 string, not less than 1 or more than 128 bytes long.

  The optional name of a target table.

Operations

- CreateIntegration action (Python: create_integration)
- ModifyIntegration action (Python: modify_integration)
- DescribeIntegrations action (Python: describe_integrations)
- DeleteIntegration action (Python: delete_integration)
- DescribeInboundIntegrations action (Python: describe_inbound_integrations)
- CreateIntegrationTableProperties action (Python: create_integration_table_properties)
- UpdateIntegrationTableProperties action (Python: update_integration_table_properties)
- GetIntegrationTableProperties action (Python: get_integration_table_properties)
- DeleteIntegrationTableProperties action (Python: delete_integration_table_properties)
- CreateIntegrationResourceProperty action (Python: create_integration_resource_property)
- UpdateIntegrationResourceProperty action (Python: update_integration_resource_property)
- GetIntegrationResourceProperty action (Python: get_integration_resource_property)
- UntagResource action (Python: untag_resource)
- ListTagsForResource action (Python: list_tags_for_resource)

CreateIntegration action (Python: create_integration)

Creates a Zero-ETL integration in the caller's account between two resources with Amazon Resource Names (ARNs): the `SourceArn` and `TargetArn`.

Request

- `IntegrationName` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  A unique name for an integration in AWS Glue.
- `SourceArn` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The ARN of the source resource for the integration.
- `TargetArn` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The ARN of the target resource for the integration.
- `Description` – UTF-8 string, not more than 1000 bytes long, matching the Custom string pattern #12.

A description of the integration.

- `DataFilter` – UTF-8 string, not less than 1 or more than 2048 bytes long.

  Selects source tables for the integration using Maxwell filter syntax.
- `KmsKeyId` – UTF-8 string, not less than 1 or more than 2048 bytes long.

  The ARN of a KMS key used for encrypting the channel.
- `AdditionalEncryptionContext` – A map array of key-value pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

  An optional set of non-secret key–value pairs that contains additional contextual information for encryption. This can only be provided if `KMSKeyId` is provided.
- `Tags` – An array of Tag objects.

  Metadata assigned to the resource consisting of a list of key-value pairs.
- `IntegrationConfig` – An IntegrationConfig object.

  The configuration settings.

Response

- `SourceArn` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The ARN of the source resource for the integration.
- `TargetArn` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The ARN of the target resource for the integration.
- `IntegrationName` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  A unique name for an integration in AWS Glue.
- `Description` – UTF-8 string, not more than 1000 bytes long, matching the Custom string pattern #12.

  A description of the integration.
- `IntegrationArn` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

The Amazon Resource Name (ARN) for the created integration.

- `KmsKeyId` – UTF-8 string, not less than 1 or more than 2048 bytes long.

  The ARN of a KMS key used for encrypting the channel.
- `AdditionalEncryptionContext` – A map array of key-value pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

  An optional set of non-secret key–value pairs that contains additional contextual information for encryption.
- `Tags` – An array of [Tag](#) objects.

  Metadata assigned to the resource consisting of a list of key-value pairs.
- `Status` – *Required:* UTF-8 string (valid values: `CREATING` | `ACTIVE` | `MODIFYING` | `FAILED` | `DELETING` | `SYNCING` | `NEEDS_ATTENTION`).

  The status of the integration being created.

  The possible statuses are:
  - CREATING: The integration is being created.
  - ACTIVE: The integration creation succeeds.
  - MODIFYING: The integration is being modified.
  - FAILED: The integration creation fails.
  - DELETING: The integration is deleted.
  - SYNCING: The integration is synchronizing.
  - NEEDS_ATTENTION: The integration needs attention, such as synchronization.
- `CreateTime` – *Required:* Timestamp.

  The time when the integration was created, in UTC.
- `Errors` – An array of [IntegrationError](#) objects.

  A list of errors associated with the integration creation.
- `DataFilter` – UTF-8 string, not less than 1 or more than 2048 bytes long.

Selects source tables for the integration using Maxwell filter syntax.

- `IntegrationConfig` – An IntegrationConfig object.

  The configuration settings.

Errors

- `ValidationException`
- `AccessDeniedException`
- `ResourceNotFoundException`
- `InternalServerException`
- `IntegrationConflictOperationFault`
- `IntegrationQuotaExceededFault`
- `KMSKeyNotAccessibleFault`
- `EntityNotFoundException`
- `InternalServiceException`
- `ConflictException`
- `ResourceNumberLimitExceededException`
- `InvalidInputException`

ModifyIntegration action (Python: modify_integration)

Modifies a Zero-ETL integration in the caller's account.

Request

- `IntegrationIdentifier` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The Amazon Resource Name (ARN) for the integration.

- `Description` – UTF-8 string, not more than 1000 bytes long, matching the Custom string pattern #12.

  A description of the integration.

- `DataFilter` – UTF-8 string, not less than 1 or more than 2048 bytes long.

  Selects source tables for the integration using Maxwell filter syntax.
- `IntegrationConfig` – An IntegrationConfig object.

  The configuration settings for the integration. Currently, only the RefreshInterval can be modified.
- `IntegrationName` – UTF-8 string, not less than 1 or more than 128 bytes long.

  A unique name for an integration in AWS Glue.

Response

- `SourceArn` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The ARN of the source for the integration.
- `TargetArn` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The ARN of the target for the integration.
- `IntegrationName` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  A unique name for an integration in AWS Glue.
- `Description` – UTF-8 string, not more than 1000 bytes long, matching the Custom string pattern #12.

  A description of the integration.
- `IntegrationArn` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The Amazon Resource Name (ARN) for the integration.
- `KmsKeyId` – UTF-8 string, not less than 1 or more than 2048 bytes long.

  The ARN of a KMS key used for encrypting the channel.
- `AdditionalEncryptionContext` – A map array of key-value pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

An optional set of non-secret key–value pairs that contains additional contextual information for encryption.

- `Tags` – An array of [Tag](#) objects.

  Metadata assigned to the resource consisting of a list of key-value pairs.
- `Status` – *Required:* UTF-8 string (valid values: `CREATING` | `ACTIVE` | `MODIFYING` | `FAILED` | `DELETING` | `SYNCING` | `NEEDS_ATTENTION`).

  The status of the integration being modified.

  The possible statuses are:
  - CREATING: The integration is being created.
  - ACTIVE: The integration creation succeeds.
  - MODIFYING: The integration is being modified.
  - FAILED: The integration creation fails.
  - DELETING: The integration is deleted.
  - SYNCING: The integration is synchronizing.
  - NEEDS_ATTENTION: The integration needs attention, such as synchronization.
- `CreateTime` – *Required:* Timestamp.

  The time when the integration was created, in UTC.
- `Errors` – An array of [IntegrationError](#) objects.

  A list of errors associated with the integration modification.
- `DataFilter` – UTF-8 string, not less than 1 or more than 2048 bytes long.

  Selects source tables for the integration using Maxwell filter syntax.
- `IntegrationConfig` – An [IntegrationConfig](#) object.

  The updated configuration settings for the integration.

Errors

- `ValidationException`
- `AccessDeniedException`

- `InternalServerException`
- `IntegrationNotFoundFault`
- `IntegrationConflictOperationFault`
- `InvalidIntegrationStateFault`
- `EntityNotFoundException`
- `InternalServiceException`
- `ConflictException`
- `InvalidStateException`
- `InvalidInputException`

DescribeIntegrations action (Python: describe_integrations)

The API is used to retrieve a list of integrations.

Request

- `IntegrationIdentifier` – UTF-8 string, not less than 1 or more than 128 bytes long.

  The Amazon Resource Name (ARN) for the integration.
- `Marker` – UTF-8 string, not less than 1 or more than 128 bytes long.

  A value that indicates the starting point for the next set of response records in a subsequent request.
- `MaxRecords` – Number (integer).

  The total number of items to return in the output.
- `Filters` – An array of IntegrationFilter objects.

  A list of key and values, to filter down the results. Supported keys are "Status", "IntegrationName", and "SourceArn". IntegrationName is limited to only one value.

Response

- `Integrations` – An array of Integration objects.

  A list of zero-ETL integrations.

- `Marker` – UTF-8 string, not less than 1 or more than 128 bytes long.

  A value that indicates the starting point for the next set of response records in a subsequent request.

Errors

- `ValidationException`
- `AccessDeniedException`
- `InternalServerException`
- `IntegrationNotFoundFault`
- `EntityNotFoundException`
- `InternalServiceException`
- `InvalidInputException`

DeleteIntegration action (Python: delete_integration)

Deletes the specified Zero-ETL integration.

Request

- `IntegrationIdentifier` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The Amazon Resource Name (ARN) for the integration.

Response

- `SourceArn` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The ARN of the source for the integration.
- `TargetArn` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The ARN of the target for the integration.
- `IntegrationName` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  A unique name for an integration in AWS Glue.

- `Description` – UTF-8 string, not more than 1000 bytes long, matching the Custom string pattern #12.

  A description of the integration.
- `IntegrationArn` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The Amazon Resource Name (ARN) for the integration.
- `KmsKeyId` – UTF-8 string, not less than 1 or more than 2048 bytes long.

  The ARN of a KMS key used for encrypting the channel.
- `AdditionalEncryptionContext` – A map array of key-value pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

  An optional set of non-secret key–value pairs that contains additional contextual information for encryption.
- `Tags` – An array of Tag objects.

  Metadata assigned to the resource consisting of a list of key-value pairs.
- `Status` – *Required:* UTF-8 string (valid values: `CREATING` | `ACTIVE` | `MODIFYING` | `FAILED` | `DELETING` | `SYNCING` | `NEEDS_ATTENTION`).

  The status of the integration being deleted.

  The possible statuses are:
  - CREATING: The integration is being created.
  - ACTIVE: The integration creation succeeds.
  - MODIFYING: The integration is being modified.
  - FAILED: The integration creation fails.
  - DELETING: The integration is deleted.
  - SYNCING: The integration is synchronizing.
  - NEEDS_ATTENTION: The integration needs attention, such as synchronization.
- `CreateTime` – *Required:* Timestamp.

  The time when the integration was created, in UTC.

- `Errors` – An array of IntegrationError objects.

  A list of errors associated with the integration.
- `DataFilter` – UTF-8 string, not less than 1 or more than 2048 bytes long.

  Selects source tables for the integration using Maxwell filter syntax.

Errors

- `ValidationException`
- `AccessDeniedException`
- `InternalServerException`
- `IntegrationNotFoundFault`
- `IntegrationConflictOperationFault`
- `InvalidIntegrationStateFault`
- `EntityNotFoundException`
- `InternalServiceException`
- `ConflictException`
- `InvalidStateException`
- `InvalidInputException`

DescribeInboundIntegrations action (Python: describe_inbound_integrations)

Returns a list of inbound integrations for the specified integration.

Request

- `IntegrationArn` – UTF-8 string, not less than 1 or more than 128 bytes long.

  The Amazon Resource Name (ARN) of the integration.
- `Marker` – UTF-8 string, not less than 1 or more than 128 bytes long.

  A token to specify where to start paginating. This is the marker from a previously truncated response.

- `MaxRecords` – Number (integer).

  The total number of items to return in the output.
- `TargetArn` – UTF-8 string, not less than 1 or more than 128 bytes long.

  The Amazon Resource Name (ARN) of the target resource in the integration.

Response

- `InboundIntegrations` – An array of InboundIntegration objects.

  A list of inbound integrations.
- `Marker` – UTF-8 string, not less than 1 or more than 128 bytes long.

  A value that indicates the starting point for the next set of response records in a subsequent request.

Errors

- `ValidationException`
- `AccessDeniedException`
- `InternalServerException`
- `IntegrationNotFoundFault`
- `TargetResourceNotFound`
- `OperationNotSupportedException`
- `EntityNotFoundException`
- `InternalServiceException`
- `InvalidInputException`

CreateIntegrationTableProperties action (Python: create_integration_table_properties)

This API is used to provide optional override properties for the the tables that need to be replicated. These properties can include properties for filtering and partitioning for the source and target tables. To set both source and target properties the same API need to be invoked with the AWS Glue connection ARN as `ResourceArn` with `SourceTableConfig`, and the AWS Glue database ARN as `ResourceArn` with `TargetTableConfig` respectively.

Request

- `ResourceArn` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The Amazon Resource Name (ARN) of the target table for which to create integration table properties. Currently, this API only supports creating integration table properties for target tables, and the provided ARN should be the ARN of the target table in the AWS Glue Data Catalog. Support for creating integration table properties for source connections (using the connection ARN) is not yet implemented and will be added in a future release.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The name of the table to be replicated.
- `SourceTableConfig` – A SourceTableConfig object.

  A structure for the source table configuration. See the `SourceTableConfig` structure to see list of supported source properties.
- `TargetTableConfig` – A TargetTableConfig object.

  A structure for the target table configuration.

Response

- *No Response parameters.*

Errors

- `ValidationException`
- `AccessDeniedException`
- `ResourceNotFoundException`
- `InternalServerException`
- `EntityNotFoundException`
- `InternalServiceException`
- `InvalidInputException`

UpdateIntegrationTableProperties action (Python: update_integration_table_properties)

This API is used to provide optional override properties for the tables that need to be replicated. These properties can include properties for filtering and partitioning for the source and target tables. To set both source and target properties the same API need to be invoked with the AWS Glue connection ARN as `ResourceArn` with `SourceTableConfig`, and the AWS Glue database ARN as `ResourceArn` with `TargetTableConfig` respectively.

The override will be reflected across all the integrations using same `ResourceArn` and source table.

Request

- `ResourceArn` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The connection ARN of the source, or the database ARN of the target.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The name of the table to be replicated.
- `SourceTableConfig` – A SourceTableConfig object.

  A structure for the source table configuration.
- `TargetTableConfig` – A TargetTableConfig object.

  A structure for the target table configuration.

Response

- *No Response parameters.*

Errors

- `ValidationException`
- `AccessDeniedException`
- `ResourceNotFoundException`
- `InternalServerException`

- `EntityNotFoundException`
- `InternalServiceException`
- `InvalidInputException`

GetIntegrationTableProperties action (Python: get_integration_table_properties)

This API is used to retrieve optional override properties for the tables that need to be replicated. These properties can include properties for filtering and partition for source and target tables.

Request

- `ResourceArn` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The Amazon Resource Name (ARN) of the target table for which to retrieve integration table properties. Currently, this API only supports retrieving properties for target tables, and the provided ARN should be the ARN of the target table in the AWS Glue Data Catalog. Support for retrieving integration table properties for source connections (using the connection ARN) is not yet implemented and will be added in a future release.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The name of the table to be replicated.

Response

- `ResourceArn` – UTF-8 string, not less than 1 or more than 128 bytes long.

  The Amazon Resource Name (ARN) of the target table for which to retrieve integration table properties. Currently, this API only supports retrieving properties for target tables, and the provided ARN should be the ARN of the target table in the AWS Glue Data Catalog. Support for retrieving integration table properties for source connections (using the connection ARN) is not yet implemented and will be added in a future release.
- `TableName` – UTF-8 string, not less than 1 or more than 128 bytes long.

  The name of the table to be replicated.
- `SourceTableConfig` – A SourceTableConfig object.

A structure for the source table configuration.

- `TargetTableConfig` – A TargetTableConfig object.

  A structure for the target table configuration.

Errors

- `ValidationException`
- `AccessDeniedException`
- `ResourceNotFoundException`
- `InternalServerException`
- `EntityNotFoundException`
- `InternalServiceException`
- `InvalidInputException`

DeleteIntegrationTableProperties action (Python: delete_integration_table_properties)

Deletes the table properties that have been created for the tables that need to be replicated.

Request

- `ResourceArn` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The connection ARN of the source, or the database ARN of the target.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The name of the table to be replicated.

Response

- *No Response parameters.*

Errors

- `ValidationException`
- `AccessDeniedException`
- `ResourceNotFoundException`
- `InternalServerException`
- `EntityNotFoundException`
- `InternalServiceException`
- `InvalidInputException`

CreateIntegrationResourceProperty action (Python: create_integration_resource_property)

This API can be used for setting up the `ResourceProperty` of the AWS Glue connection (for the source) or AWS Glue database ARN (for the target). These properties can include the role to access the connection or database. To set both source and target properties the same API needs to be invoked with the AWS Glue connection ARN as `ResourceArn` with `SourceProcessingProperties` and the AWS Glue database ARN as `ResourceArn` with `TargetProcessingProperties` respectively.

Request

- `ResourceArn` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The connection ARN of the source, or the database ARN of the target.
- `SourceProcessingProperties` – A SourceProcessingProperties object.

  The resource properties associated with the integration source.
- `TargetProcessingProperties` – A TargetProcessingProperties object.

  The resource properties associated with the integration target.

Response

- `ResourceArn` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The connection ARN of the source, or the database ARN of the target.

- `SourceProcessingProperties` – A SourceProcessingProperties object.

  The resource properties associated with the integration source.
- `TargetProcessingProperties` – A TargetProcessingProperties object.

  The resource properties associated with the integration target.

Errors

- `ValidationException`
- `AccessDeniedException`
- `ConflictException`
- `InternalServerException`
- `ResourceNotFoundException`
- `EntityNotFoundException`
- `InternalServiceException`
- `InvalidInputException`

UpdateIntegrationResourceProperty action (Python: update_integration_resource_property)

This API can be used for updating the `ResourceProperty` of the AWS Glue connection (for the source) or AWS Glue database ARN (for the target). These properties can include the role to access the connection or database. Since the same resource can be used across multiple integrations, updating resource properties will impact all the integrations using it.

Request

- `ResourceArn` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The connection ARN of the source, or the database ARN of the target.
- `SourceProcessingProperties` – A SourceProcessingProperties object.

  The resource properties associated with the integration source.
- `TargetProcessingProperties` – A TargetProcessingProperties object.

The resource properties associated with the integration target.

Response

- `ResourceArn` – UTF-8 string, not less than 1 or more than 128 bytes long.

  The connection ARN of the source, or the database ARN of the target.
- `SourceProcessingProperties` – A SourceProcessingProperties object.

  The resource properties associated with the integration source.
- `TargetProcessingProperties` – A TargetProcessingProperties object.

  The resource properties associated with the integration target.

Errors

- `ValidationException`
- `AccessDeniedException`
- `InternalServerException`
- `ResourceNotFoundException`
- `EntityNotFoundException`
- `InternalServiceException`
- `InvalidInputException`

GetIntegrationResourceProperty action (Python: get_integration_resource_property)

This API is used for fetching the `ResourceProperty` of the AWS Glue connection (for the source) or AWS Glue database ARN (for the target)

Request

- `ResourceArn` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long.

  The connection ARN of the source, or the database ARN of the target.

Response

- `ResourceArn` – UTF-8 string, not less than 1 or more than 128 bytes long.

  The connection ARN of the source, or the database ARN of the target.
- `SourceProcessingProperties` – A SourceProcessingProperties object.

  The resource properties associated with the integration source.
- `TargetProcessingProperties` – A TargetProcessingProperties object.

  The resource properties associated with the integration target.

Errors

- `ValidationException`
- `AccessDeniedException`
- `InternalServerException`
- `ResourceNotFoundException`
- `EntityNotFoundException`
- `InternalServiceException`
- `InvalidInputException`

UntagResource action (Python: untag_resource)

Removes the specified tags from an integration resource.

Request

- `ResourceArn` – *Required:* UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

  The Amazon Resource Name (ARN) for the integration resource.
- `TagsToRemove` – *Required:* An array of UTF-8 strings, not more than 50 strings.

  A list of metadata tags to be removed from the resource.

Response

- *No Response parameters.*

Errors

- `ResourceNotFoundException`

ListTagsForResource action (Python: list_tags_for_resource)

Lists the metadata tags assigned to the specified resource.

Request

- `ResourceARN` – *Required:* UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

  The resource ARN for the resource.

Response

- `Tags` – An array of Tag objects, not more than 10 structures.

  A list of tags.

Errors

- `ResourceNotFoundException`

Exceptions

- ResourceNotFoundException structure
- InternalServerException structure
- IntegrationAlreadyExistsFault structure
- IntegrationConflictOperationFault structure
- IntegrationQuotaExceededFault structure
- KMSKeyNotAccessibleFault structure

- IntegrationNotFoundFault structure
- TargetResourceNotFound structure
- InvalidIntegrationStateFault structure

ResourceNotFoundException structure

The resource could not be found.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

InternalServerException structure

An internal server error occurred.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

IntegrationAlreadyExistsFault structure

The specified integration already exists.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

IntegrationConflictOperationFault structure

The requested operation conflicts with another operation.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

IntegrationQuotaExceededFault structure

The data processed through your integration exceeded your quota.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

KMSKeyNotAccessibleFault structure

The KMS key specified is not accessible.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

IntegrationNotFoundFault structure

The specified integration could not be found.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

TargetResourceNotFound structure

The target resource could not be found.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

InvalidIntegrationStateFault structure

The integration is in an invalid state.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

Interactive sessions API

The interactive sessions API describes the AWS Glue API related to using AWS Glue interactive sessions to build and test extract, transform, and load (ETL) scripts for data integration.

Data types

- Session structure
- SessionCommand structure
- Statement structure
- StatementOutput structure
- StatementOutputData structure
- ConnectionsList structure

Session structure

The period in which a remote Spark runtime environment is running.

Fields

- `Id` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The ID of the session.

- `CreatedOn` – Timestamp.

  The time and date when the session was created.
- `Status` – UTF-8 string (valid values: `PROVISIONING` | `READY` | `FAILED` | `TIMEOUT` | `STOPPING` | `STOPPED`).

  The session status.
- `ErrorMessage` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  The error message displayed during the session.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  The description of the session.
- `Role` – UTF-8 string, not less than 20 or more than 2048 bytes long, matching the Custom string pattern #30.

  The name or Amazon Resource Name (ARN) of the IAM role associated with the Session.
- `Command` – A SessionCommand object.

  The command object.See SessionCommand.
- `DefaultArguments` – A map array of key-value pairs, not more than 75 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #31.

  Each value is a UTF-8 string, not more than 4096 bytes long, matching the URI address multi-line string pattern.

  A map array of key-value pairs. Max is 75 pairs.
- `Connections` – A ConnectionsList object.

  The number of connections used for the session.
- `Progress` – Number (double).

  The code execution progress of the session.

- `MaxCapacity` – Number (double).

  The number of AWS Glue data processing units (DPUs) that can be allocated when the job runs. A DPU is a relative measure of processing power that consists of 4 vCPUs of compute capacity and 16 GB memory.
- `SecurityConfiguration` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the SecurityConfiguration structure to be used with the session.
- `GlueVersion` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #47.

  The AWS Glue version determines the versions of Apache Spark and Python that AWS Glue supports. The GlueVersion must be greater than 2.0.
- `DataAccessId` – UTF-8 string, not less than 1 or more than 36 bytes long.

  The data access ID of the session.
- `PartitionId` – UTF-8 string, not less than 1 or more than 36 bytes long.

  The partition ID of the sesion.
- `NumberOfWorkers` – Number (integer).

  The number of workers of a defined `WorkerType` to use for the session.
- `WorkerType` – UTF-8 string (valid values: `Standard=""` | `G.1X=""` | `G.2X=""` | `G.025X=""` | `G.4X=""` | `G.8X=""` | `Z.2X=""`).

  The type of predefined worker that is allocated when a session runs. Accepts a value of `G.1X`, `G.2X`, `G.4X`, or `G.8X` for Spark sessions. Accepts the value `Z.2X` for Ray sessions.
- `CompletedOn` – Timestamp.

  The date and time that this session is completed.
- `ExecutionTime` – Number (double).

  The total time the session ran for.
- `DPUSeconds` – Number (double).

  The DPUs consumed by the session (formula: ExecutionTime * MaxCapacity).
- `IdleTimeout` – Number (integer).

The number of minutes when idle before the session times out.

- `ProfileName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of an AWS Glue usage profile associated with the session.

SessionCommand structure

The `SessionCommand` that runs the job.

Fields

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Specifies the name of the SessionCommand. Can be 'glueetl' or 'gluestreaming'.
- `PythonVersion` – UTF-8 string, matching the Custom string pattern #48.

  Specifies the Python version. The Python version indicates the version supported for jobs of type Spark.

Statement structure

The statement or request for a particular action to occur in a session.

Fields

- `Id` – Number (integer).

  The ID of the statement.
- `Code` – UTF-8 string.

  The execution code of the statement.
- `State` – UTF-8 string (valid values: `WAITING` | `RUNNING` | `AVAILABLE` | `CANCELLING` | `CANCELLED` | `ERROR`).

  The state while request is actioned.
- `Output` – A StatementOutput object.

The output in JSON.

- `Progress` – Number (double).

  The code execution progress.

- `StartedOn` – Number (long).

  The unix time and date that the job definition was started.

- `CompletedOn` – Number (long).

  The unix time and date that the job definition was completed.

StatementOutput structure

The code execution output in JSON format.

Fields

- `Data` – A StatementOutputData object.

  The code execution output.

- `ExecutionCount` – Number (integer).

  The execution count of the output.

- `Status` – UTF-8 string (valid values: `WAITING` | `RUNNING` | `AVAILABLE` | `CANCELLING` | `CANCELLED` | `ERROR`).

  The status of the code execution output.

- `ErrorName` – UTF-8 string.

  The name of the error in the output.

- `ErrorValue` – UTF-8 string.

  The error value of the output.

- `Traceback` – An array of UTF-8 strings.

  The traceback of the output.

StatementOutputData structure

The code execution output in JSON format.

Fields

- `TextPlain` – UTF-8 string.

  The code execution output in text format.

ConnectionsList structure

Specifies the connections used by a job.

Fields

- `Connections` – An array of UTF-8 strings, not more than 1000 strings.

  A list of connections used by the job.

Operations

- CreateSession action (Python: create_session)
- StopSession action (Python: stop_session)
- DeleteSession action (Python: delete_session)
- GetSession action (Python: get_session)
- ListSessions action (Python: list_sessions)
- RunStatement action (Python: run_statement)
- CancelStatement action (Python: cancel_statement)
- GetStatement action (Python: get_statement)
- ListStatements action (Python: list_statements)
- GetGlueIdentityCenterConfiguration action (Python: get_glue_identity_center_configuration)
- UpdateGlueIdentityCenterConfiguration action (Python: update_glue_identity_center_configuration)
- CreateGlueIdentityCenterConfiguration action (Python: create_glue_identity_center_configuration)

- DeleteGlueIdentityCenterConfiguration action (Python: delete_glue_identity_center_configuration)

CreateSession action (Python: create_session)

Creates a new session.

Request

Request to create a new session.

- `Id` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the session request.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  The description of the session.
- `Role` – *Required:* UTF-8 string, not less than 20 or more than 2048 bytes long, matching the Custom string pattern #30.

  The IAM Role ARN
- `Command` – *Required:* A SessionCommand object.

  The `SessionCommand` that runs the job.
- `Timeout` – Number (integer), at least 1.

  The number of minutes before session times out. Default for Spark ETL jobs is 48 hours (2880 minutes). Consult the documentation for other job types.
- `IdleTimeout` – Number (integer), at least 1.

  The number of minutes when idle before session times out. Default for Spark ETL jobs is value of Timeout. Consult the documentation for other job types.
- `DefaultArguments` – A map array of key-value pairs, not more than 75 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #31.

Each value is a UTF-8 string, not more than 4096 bytes long, matching the URI address multi-line string pattern.

A map array of key-value pairs. Max is 75 pairs.

- `Connections` – A ConnectionsList object.

  The number of connections to use for the session.
- `MaxCapacity` – Number (double).

  The number of AWS Glue data processing units (DPUs) that can be allocated when the job runs. A DPU is a relative measure of processing power that consists of 4 vCPUs of compute capacity and 16 GB memory.
- `NumberOfWorkers` – Number (integer).

  The number of workers of a defined `WorkerType` to use for the session.
- `WorkerType` – UTF-8 string (valid values: `Standard=""` | `G.1X=""` | `G.2X=""` | `G.025X=""` | `G.4X=""` | `G.8X=""` | `Z.2X=""`).

  The type of predefined worker that is allocated when a job runs. Accepts a value of G.1X, G.2X, G.4X, or G.8X for Spark jobs. Accepts the value Z.2X for Ray notebooks.
  - For the `G.1X` worker type, each worker maps to 1 DPU (4 vCPUs, 16 GB of memory) with 94GB disk, and provides 1 executor per worker. We recommend this worker type for workloads such as data transforms, joins, and queries, to offers a scalable and cost effective way to run most jobs.
  - For the `G.2X` worker type, each worker maps to 2 DPU (8 vCPUs, 32 GB of memory) with 138GB disk, and provides 1 executor per worker. We recommend this worker type for workloads such as data transforms, joins, and queries, to offers a scalable and cost effective way to run most jobs.
  - For the `G.4X` worker type, each worker maps to 4 DPU (16 vCPUs, 64 GB of memory) with 256GB disk, and provides 1 executor per worker. We recommend this worker type for jobs whose workloads contain your most demanding transforms, aggregations, joins, and queries. This worker type is available only for AWS Glue version 3.0 or later Spark ETL jobs in the following AWS Regions: US East (Ohio), US East (N. Virginia), US West (Oregon), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Canada (Central), Europe (Frankfurt), Europe (Ireland), and Europe (Stockholm).

  - For the `G.8X` worker type, each worker maps to 8 DPU (32 vCPUs, 128 GB of memory) with 512GB disk, and provides 1 executor per worker. We recommend this worker type for jobs whose workloads contain your most demanding transforms, aggregations, joins, and queries. This worker type is available only for AWS Glue version 3.0 or later Spark ETL jobs, in the same AWS Regions as supported for the `G.4X` worker type.
  - For the `Z.2X` worker type, each worker maps to 2 M-DPU (8vCPUs, 64 GB of memory) with 128 GB disk, and provides up to 8 Ray workers based on the autoscaler.
- `SecurityConfiguration` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the SecurityConfiguration structure to be used with the session
- `GlueVersion` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #47.

  The AWS Glue version determines the versions of Apache Spark and Python that AWS Glue supports. The GlueVersion must be greater than 2.0.
- `DataAccessId` – UTF-8 string, not less than 1 or more than 36 bytes long.

  The data access ID of the session.
- `PartitionId` – UTF-8 string, not less than 1 or more than 36 bytes long.

  The partition ID of the session.
- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  The map of key value pairs (tags) belonging to the session.
- `RequestOrigin` – UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #31.

  The origin of the request.
- `ProfileName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of an AWS Glue usage profile associated with the session.

Response

- `Session` – A Session object.

  Returns the session object in the response.

Errors

- `AccessDeniedException`
- `IdempotentParameterMismatchException`
- `InternalServiceException`
- `OperationTimeoutException`
- `InvalidInputException`
- `ValidationException`
- `AlreadyExistsException`
- `ResourceNumberLimitExceededException`

StopSession action (Python: stop_session)

Stops the session.

Request

- `Id` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the session to be stopped.
- `RequestOrigin` – UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #31.

  The origin of the request.

Response

- `Id` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

Returns the Id of the stopped session.

Errors

- `AccessDeniedException`
- `InternalServiceException`
- `OperationTimeoutException`
- `InvalidInputException`
- `IllegalSessionStateException`
- `ConcurrentModificationException`

DeleteSession action (Python: delete_session)

Deletes the session.

Request

- `Id` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the session to be deleted.
- `RequestOrigin` – UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #31.

  The name of the origin of the delete session request.

Response

- `Id` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Returns the ID of the deleted session.

Errors

- `AccessDeniedException`

- `InternalServiceException`
- `OperationTimeoutException`
- `InvalidInputException`
- `IllegalSessionStateException`
- `ConcurrentModificationException`

GetSession action (Python: get_session)

Retrieves the session.

Request

- `Id` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the session.
- `RequestOrigin` – UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #31.

  The origin of the request.

Response

- `Session` – A Session object.

  The session object is returned in the response.

Errors

- `AccessDeniedException`
- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`
- `InvalidInputException`

ListSessions action (Python: list_sessions)

Retrieve a list of sessions.

Request

- `NextToken` – UTF-8 string, not more than 400000 bytes long.

  The token for the next set of results, or null if there are no more result.
- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum number of results.
- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  Tags belonging to the session.
- `RequestOrigin` – UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #31.

  The origin of the request.

Response

- `Ids` – An array of UTF-8 strings.

  Returns the ID of the session.
- `Sessions` – An array of Session objects.

  Returns the session object.
- `NextToken` – UTF-8 string, not more than 400000 bytes long.

  The token for the next set of results, or null if there are no more result.

Errors

- `AccessDeniedException`

- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`

RunStatement action (Python: run_statement)

Executes the statement.

Request

- `SessionId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Session Id of the statement to be run.
- `Code` – *Required:* UTF-8 string, not more than 68000 bytes long.

  The statement code to be run.
- `RequestOrigin` – UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #31.

  The origin of the request.

Response

- `Id` – Number (integer).

  Returns the Id of the statement that was run.

Errors

- `EntityNotFoundException`
- `AccessDeniedException`
- `InternalServiceException`
- `OperationTimeoutException`
- `InvalidInputException`
- `ValidationException`

- `ResourceNumberLimitExceededException`
- `IllegalSessionStateException`

CancelStatement action (Python: cancel_statement)

Cancels the statement.

Request

- `SessionId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Session ID of the statement to be cancelled.
- `Id` – *Required:* Number (integer).

  The ID of the statement to be cancelled.
- `RequestOrigin` – UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #31.

  The origin of the request to cancel the statement.

Response

- *No Response parameters.*

Errors

- `AccessDeniedException`
- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`
- `InvalidInputException`
- `IllegalSessionStateException`

GetStatement action (Python: get_statement)

Retrieves the statement.

Request

- `SessionId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Session ID of the statement.
- `Id` – *Required:* Number (integer).

  The Id of the statement.
- `RequestOrigin` – UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #31.

  The origin of the request.

Response

- `Statement` – A Statement object.

  Returns the statement.

Errors

- `AccessDeniedException`
- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`
- `InvalidInputException`
- `IllegalSessionStateException`

ListStatements action (Python: list_statements)

Lists statements for the session.

Request

- `SessionId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Session ID of the statements.
- `RequestOrigin` – UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #31.

  The origin of the request to list statements.
- `NextToken` – UTF-8 string, not more than 400000 bytes long.

  A continuation token, if this is a continuation call.

Response

- `Statements` – An array of Statement objects.

  Returns the list of statements.
- `NextToken` – UTF-8 string, not more than 400000 bytes long.

  A continuation token, if not all statements have yet been returned.

Errors

- `AccessDeniedException`
- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`
- `InvalidInputException`
- `IllegalSessionStateException`

GetGlueIdentityCenterConfiguration action (Python: get_glue_identity_center_configuration)

Retrieves the current AWS Glue Identity Center configuration details, including the associated Identity Center instance and application information.

Request

- *No Request parameters.*

Response

Response containing the AWS Glue Identity Center configuration details.

- `ApplicationArn` – UTF-8 string, not less than 10 or more than 1224 bytes long.

  The Amazon Resource Name (ARN) of the Identity Center application associated with the AWS Glue configuration.
- `InstanceArn` – UTF-8 string, not less than 10 or more than 1224 bytes long.

  The Amazon Resource Name (ARN) of the Identity Center instance associated with the AWS Glue configuration.
- `Scopes` – An array of UTF-8 strings.

  A list of Identity Center scopes that define the permissions and access levels for the AWS Glue configuration.

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`
- `AccessDeniedException`
- `ConcurrentModificationException`

UpdateGlueIdentityCenterConfiguration action (Python: update_glue_identity_center_configuration)

Updates the existing AWS Glue Identity Center configuration, allowing modification of scopes and permissions for the integration.

Request

Request to update an existing AWS Glue Identity Center configuration.

- `Scopes` – An array of UTF-8 strings, not less than 1 or more than 50 strings.

  A list of Identity Center scopes that define the updated permissions and access levels for the AWS Glue configuration.

Response

- *No Response parameters.*

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`
- `AccessDeniedException`
- `ConcurrentModificationException`

CreateGlueIdentityCenterConfiguration action (Python: create_glue_identity_center_configuration)

Creates a new AWS Glue Identity Center configuration to enable integration between AWS Glue and AWS IAM Identity Center for authentication and authorization.

Request

Request to create a new AWS Glue Identity Center configuration.

- `InstanceArn` – *Required:* UTF-8 string, not less than 10 or more than 1224 bytes long.

  The Amazon Resource Name (ARN) of the Identity Center instance to be associated with the AWS Glue configuration.
- `Scopes` – An array of UTF-8 strings, not less than 1 or more than 50 strings.

  A list of Identity Center scopes that define the permissions and access levels for the AWS Glue configuration.

Response

Response from creating a new AWS Glue Identity Center configuration.

- `ApplicationArn` – UTF-8 string, not less than 10 or more than 1224 bytes long.

  The Amazon Resource Name (ARN) of the Identity Center application that was created for the AWS Glue configuration.

Errors

- `InvalidInputException`
- `AlreadyExistsException`
- `InternalServiceException`
- `OperationTimeoutException`
- `AccessDeniedException`
- `ConcurrentModificationException`

DeleteGlueIdentityCenterConfiguration action (Python: delete_glue_identity_center_configuration)

Deletes the existing AWS Glue Identity Center configuration, removing the integration between AWS Glue and AWS IAM Identity Center.

Request

- *No Request parameters.*

Response

- *No Response parameters.*

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`
- `AccessDeniedException`
- `ConcurrentModificationException`

Development endpoints API

The Development endpoints API describes the AWS Glue API related to testing using a custom DevEndpoint.

Data types

- [DevEndpoint structure]
- [DevEndpointCustomLibraries structure]

DevEndpoint structure

A development endpoint where a developer can remotely debug extract, transform, and load (ETL) scripts.

Fields

- `EndpointName` – UTF-8 string.

  The name of the `DevEndpoint`.
- `RoleArn` – UTF-8 string, matching the AWS IAM ARN string pattern.

  The Amazon Resource Name (ARN) of the IAM role used in this `DevEndpoint`.

- `SecurityGroupIds` – An array of UTF-8 strings.

  A list of security group identifiers used in this `DevEndpoint`.
- `SubnetId` – UTF-8 string.

  The subnet ID for this `DevEndpoint`.
- `YarnEndpointAddress` – UTF-8 string.

  The YARN endpoint address used by this `DevEndpoint`.
- `PrivateAddress` – UTF-8 string.

  A private IP address to access the `DevEndpoint` within a VPC if the `DevEndpoint` is created within one. The `PrivateAddress` field is present only when you create the `DevEndpoint` within your VPC.
- `ZeppelinRemoteSparkInterpreterPort` – Number (integer).

  The Apache Zeppelin port for the remote Apache Spark interpreter.
- `PublicAddress` – UTF-8 string.

  The public IP address used by this `DevEndpoint`. The `PublicAddress` field is present only when you create a non-virtual private cloud (VPC) `DevEndpoint`.
- `Status` – UTF-8 string.

  The current status of this `DevEndpoint`.
- `WorkerType` – UTF-8 string (valid values: `Standard=""` | `G.1X=""` | `G.2X=""` | `G.025X=""` | `G.4X=""` | `G.8X=""` | `Z.2X=""`).

  The type of predefined worker that is allocated to the development endpoint. Accepts a value of Standard, G.1X, or G.2X.
  - For the `Standard` worker type, each worker provides 4 vCPU, 16 GB of memory and a 50GB disk, and 2 executors per worker.
  - For the `G.1X` worker type, each worker maps to 1 DPU (4 vCPU, 16 GB of memory, 64 GB disk), and provides 1 executor per worker. We recommend this worker type for memory-intensive jobs.
  - For the `G.2X` worker type, each worker maps to 2 DPU (8 vCPU, 32 GB of memory, 128 GB disk), and provides 1 executor per worker. We recommend this worker type for memory-intensive jobs.

Known issue: when a development endpoint is created with the `G.2X` `WorkerType` configuration, the Spark drivers for the development endpoint will run on 4 vCPU, 16 GB of memory, and a 64 GB disk.

- `GlueVersion` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #47.

  Glue version determines the versions of Apache Spark and Python that AWS Glue supports. The Python version indicates the version supported for running your ETL scripts on development endpoints.

  For more information about the available AWS Glue versions and corresponding Spark and Python versions, see Glue version in the developer guide.

  Development endpoints that are created without specifying a Glue version default to Glue 0.9.

  You can specify a version of Python support for development endpoints by using the `Arguments` parameter in the `CreateDevEndpoint` or `UpdateDevEndpoint` APIs. If no arguments are provided, the version defaults to Python 2.
- `NumberOfWorkers` – Number (integer).

  The number of workers of a defined `workerType` that are allocated to the development endpoint.

  The maximum number of workers you can define are 299 for `G.1X`, and 149 for `G.2X`.
- `NumberOfNodes` – Number (integer).

  The number of AWS Glue Data Processing Units (DPUs) allocated to this `DevEndpoint`.
- `AvailabilityZone` – UTF-8 string.

  The AWS Availability Zone where this `DevEndpoint` is located.
- `VpcId` – UTF-8 string.

  The ID of the virtual private cloud (VPC) used by this `DevEndpoint`.
- `ExtraPythonLibsS3Path` – UTF-8 string.

  The paths to one or more Python libraries in an Amazon S3 bucket that should be loaded in your `DevEndpoint`. Multiple values must be complete paths separated by a comma.

> **Note**
>
> You can only use pure Python libraries with a `DevEndpoint`. Libraries that rely on C extensions, such as the pandas Python data analysis library, are not currently supported.

- `ExtraJarsS3Path` – UTF-8 string.

  The path to one or more Java `.jar` files in an S3 bucket that should be loaded in your `DevEndpoint`.

  > **Note**
  >
  > You can only use pure Java/Scala libraries with a `DevEndpoint`.

- `FailureReason` – UTF-8 string.

  The reason for a current failure in this `DevEndpoint`.
- `LastUpdateStatus` – UTF-8 string.

  The status of the last update.
- `CreatedTimestamp` – Timestamp.

  The point in time at which this DevEndpoint was created.
- `LastModifiedTimestamp` – Timestamp.

  The point in time at which this `DevEndpoint` was last modified.
- `PublicKey` – UTF-8 string.

  The public key to be used by this `DevEndpoint` for authentication. This attribute is provided for backward compatibility because the recommended attribute to use is public keys.
- `PublicKeys` – An array of UTF-8 strings, not more than 5 strings.

  A list of public keys to be used by the `DevEndpoints` for authentication. Using this attribute is preferred over a single public key because the public keys allow you to have a different private key per client.

> **Note**
>
> If you previously created an endpoint with a public key, you must remove that key to be able to set a list of public keys. Call the `UpdateDevEndpoint` API operation with the public key content in the `deletePublicKeys` attribute, and the list of new keys in the `addPublicKeys` attribute.

- `SecurityConfiguration` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the `SecurityConfiguration` structure to be used with this `DevEndpoint`.
- `Arguments` – A map array of key-value pairs, not more than 100 pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

  A map of arguments used to configure the `DevEndpoint`.

  Valid arguments are:

  - `"--enable-glue-datacatalog": ""`

  You can specify a version of Python support for development endpoints by using the `Arguments` parameter in the `CreateDevEndpoint` or `UpdateDevEndpoint` APIs. If no arguments are provided, the version defaults to Python 2.

DevEndpointCustomLibraries structure

Custom libraries to be loaded into a development endpoint.

Fields

- `ExtraPythonLibsS3Path` – UTF-8 string.

  The paths to one or more Python libraries in an Amazon Simple Storage Service (Amazon S3) bucket that should be loaded in your `DevEndpoint`. Multiple values must be complete paths separated by a comma.

> **Note**
>
> You can only use pure Python libraries with a `DevEndpoint`. Libraries that rely on C extensions, such as the pandas Python data analysis library, are not currently supported.

- `ExtraJarsS3Path` – UTF-8 string.

  The path to one or more Java `.jar` files in an S3 bucket that should be loaded in your `DevEndpoint`.

> **Note**
>
> You can only use pure Java/Scala libraries with a `DevEndpoint`.

Operations

- CreateDevEndpoint action (Python: create_dev_endpoint)
- UpdateDevEndpoint action (Python: update_dev_endpoint)
- DeleteDevEndpoint action (Python: delete_dev_endpoint)
- GetDevEndpoint action (Python: get_dev_endpoint)
- GetDevEndpoints action (Python: get_dev_endpoints)
- BatchGetDevEndpoints action (Python: batch_get_dev_endpoints)
- ListDevEndpoints action (Python: list_dev_endpoints)

CreateDevEndpoint action (Python: create_dev_endpoint)

Creates a new development endpoint.

Request

- `EndpointName` – *Required:* UTF-8 string.

  The name to be assigned to the new `DevEndpoint`.

- `RoleArn` – *Required:* UTF-8 string, matching the AWS IAM ARN string pattern.

The IAM role for the `DevEndpoint`.

- `SecurityGroupIds` – An array of UTF-8 strings.

  Security group IDs for the security groups to be used by the new `DevEndpoint`.
- `SubnetId` – UTF-8 string.

  The subnet ID for the new `DevEndpoint` to use.
- `PublicKey` – UTF-8 string.

  The public key to be used by this `DevEndpoint` for authentication. This attribute is provided for backward compatibility because the recommended attribute to use is public keys.
- `PublicKeys` – An array of UTF-8 strings, not more than 5 strings.

  A list of public keys to be used by the development endpoints for authentication. The use of this attribute is preferred over a single public key because the public keys allow you to have a different private key per client.

  > **Note**
  >
  > If you previously created an endpoint with a public key, you must remove that key to be able to set a list of public keys. Call the `UpdateDevEndpoint` API with the public key content in the `deletePublicKeys` attribute, and the list of new keys in the `addPublicKeys` attribute.
- `NumberOfNodes` – Number (integer).

  The number of AWS Glue Data Processing Units (DPUs) to allocate to this `DevEndpoint`.
- `WorkerType` – UTF-8 string (valid values: `Standard=""` | `G.1X=""` | `G.2X=""` | `G.025X=""` | `G.4X=""` | `G.8X=""` | `Z.2X=""`).

  The type of predefined worker that is allocated to the development endpoint. Accepts a value of Standard, G.1X, or G.2X.
  - For the `Standard` worker type, each worker provides 4 vCPU, 16 GB of memory and a 50GB disk, and 2 executors per worker.
  - For the `G.1X` worker type, each worker maps to 1 DPU (4 vCPU, 16 GB of memory, 64 GB disk), and provides 1 executor per worker. We recommend this worker type for memory-intensive jobs.

  - For the `G.2X` worker type, each worker maps to 2 DPU (8 vCPU, 32 GB of memory, 128 GB disk), and provides 1 executor per worker. We recommend this worker type for memory-intensive jobs.

  Known issue: when a development endpoint is created with the `G.2X` `WorkerType` configuration, the Spark drivers for the development endpoint will run on 4 vCPU, 16 GB of memory, and a 64 GB disk.

- `GlueVersion` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #47.

  Glue version determines the versions of Apache Spark and Python that AWS Glue supports. The Python version indicates the version supported for running your ETL scripts on development endpoints.

  For more information about the available AWS Glue versions and corresponding Spark and Python versions, see Glue version in the developer guide.

  Development endpoints that are created without specifying a Glue version default to Glue 0.9.

  You can specify a version of Python support for development endpoints by using the `Arguments` parameter in the `CreateDevEndpoint` or `UpdateDevEndpoint` APIs. If no arguments are provided, the version defaults to Python 2.
- `NumberOfWorkers` – Number (integer).

  The number of workers of a defined `workerType` that are allocated to the development endpoint.

  The maximum number of workers you can define are 299 for `G.1X`, and 149 for `G.2X`.
- `ExtraPythonLibsS3Path` – UTF-8 string.

  The paths to one or more Python libraries in an Amazon S3 bucket that should be loaded in your `DevEndpoint`. Multiple values must be complete paths separated by a comma.

  > **Note**
  >
  > You can only use pure Python libraries with a `DevEndpoint`. Libraries that rely on C extensions, such as the pandas Python data analysis library, are not yet supported.
- `ExtraJarsS3Path` – UTF-8 string.

The path to one or more Java `.jar` files in an S3 bucket that should be loaded in your `DevEndpoint`.

- `SecurityConfiguration` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the `SecurityConfiguration` structure to be used with this `DevEndpoint`.
- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  The tags to use with this DevEndpoint. You may use tags to limit access to the DevEndpoint. For more information about tags in AWS Glue, see AWS Tags in AWS Glue in the developer guide.
- `Arguments` – A map array of key-value pairs, not more than 100 pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

  A map of arguments used to configure the `DevEndpoint`.

Response

- `EndpointName` – UTF-8 string.

  The name assigned to the new `DevEndpoint`.
- `Status` – UTF-8 string.

  The current status of the new `DevEndpoint`.
- `SecurityGroupIds` – An array of UTF-8 strings.

  The security groups assigned to the new `DevEndpoint`.
- `SubnetId` – UTF-8 string.

  The subnet ID assigned to the new `DevEndpoint`.
- `RoleArn` – UTF-8 string, matching the AWS IAM ARN string pattern.

The Amazon Resource Name (ARN) of the role assigned to the new `DevEndpoint`.

- `YarnEndpointAddress` – UTF-8 string.

  The address of the YARN endpoint used by this `DevEndpoint`.
- `ZeppelinRemoteSparkInterpreterPort` – Number (integer).

  The Apache Zeppelin port for the remote Apache Spark interpreter.
- `NumberOfNodes` – Number (integer).

  The number of AWS Glue Data Processing Units (DPUs) allocated to this DevEndpoint.
- `WorkerType` – UTF-8 string (valid values: `Standard=""` | `G.1X=""` | `G.2X=""` | `G.025X=""` | `G.4X=""` | `G.8X=""` | `Z.2X=""`).

  The type of predefined worker that is allocated to the development endpoint. May be a value of Standard, G.1X, or G.2X.
- `GlueVersion` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #47.

  Glue version determines the versions of Apache Spark and Python that AWS Glue supports. The Python version indicates the version supported for running your ETL scripts on development endpoints.

  For more information about the available AWS Glue versions and corresponding Spark and Python versions, see Glue version in the developer guide.
- `NumberOfWorkers` – Number (integer).

  The number of workers of a defined `workerType` that are allocated to the development endpoint.
- `AvailabilityZone` – UTF-8 string.

  The AWS Availability Zone where this `DevEndpoint` is located.
- `VpcId` – UTF-8 string.

  The ID of the virtual private cloud (VPC) used by this `DevEndpoint`.
- `ExtraPythonLibsS3Path` – UTF-8 string.

The paths to one or more Python libraries in an S3 bucket that will be loaded in your `DevEndpoint`.

- `ExtraJarsS3Path` – UTF-8 string.

  Path to one or more Java `.jar` files in an S3 bucket that will be loaded in your `DevEndpoint`.
- `FailureReason` – UTF-8 string.

  The reason for a current failure in this `DevEndpoint`.
- `SecurityConfiguration` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the `SecurityConfiguration` structure being used with this `DevEndpoint`.
- `CreatedTimestamp` – Timestamp.

  The point in time at which this `DevEndpoint` was created.
- `Arguments` – A map array of key-value pairs, not more than 100 pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

  The map of arguments used to configure this `DevEndpoint`.

  Valid arguments are:

  - `"--enable-glue-datacatalog": ""`

  You can specify a version of Python support for development endpoints by using the `Arguments` parameter in the `CreateDevEndpoint` or `UpdateDevEndpoint` APIs. If no arguments are provided, the version defaults to Python 2.

Errors

- `AccessDeniedException`
- `AlreadyExistsException`
- `IdempotentParameterMismatchException`
- `InternalServiceException`
- `OperationTimeoutException`

- `InvalidInputException`
- `ValidationException`
- `ResourceNumberLimitExceededException`

UpdateDevEndpoint action (Python: update_dev_endpoint)

Updates a specified development endpoint.

Request

- `EndpointName` – *Required:* UTF-8 string.

  The name of the `DevEndpoint` to be updated.
- `PublicKey` – UTF-8 string.

  The public key for the `DevEndpoint` to use.
- `AddPublicKeys` – An array of UTF-8 strings, not more than 5 strings.

  The list of public keys for the `DevEndpoint` to use.
- `DeletePublicKeys` – An array of UTF-8 strings, not more than 5 strings.

  The list of public keys to be deleted from the `DevEndpoint`.
- `CustomLibraries` – A DevEndpointCustomLibraries object.

  Custom Python or Java libraries to be loaded in the `DevEndpoint`.
- `UpdateEtlLibraries` – Boolean.

  `True` if the list of custom libraries to be loaded in the development endpoint needs to be updated, or `False` if otherwise.
- `DeleteArguments` – An array of UTF-8 strings.

  The list of argument keys to be deleted from the map of arguments used to configure the `DevEndpoint`.
- `AddArguments` – A map array of key-value pairs, not more than 100 pairs.

  Each key is a UTF-8 string.

  Each value is a UTF-8 string.

The map of arguments to add the map of arguments used to configure the `DevEndpoint`.

Valid arguments are:

- `"--enable-glue-datacatalog": ""`

You can specify a version of Python support for development endpoints by using the `Arguments` parameter in the `CreateDevEndpoint` or `UpdateDevEndpoint` APIs. If no arguments are provided, the version defaults to Python 2.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`
- `InvalidInputException`
- `ValidationException`

DeleteDevEndpoint action (Python: delete_dev_endpoint)

Deletes a specified development endpoint.

Request

- `EndpointName` – *Required:* UTF-8 string.

  The name of the `DevEndpoint`.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`
- `InvalidInputException`

GetDevEndpoint action (Python: get_dev_endpoint)

Retrieves information about a specified development endpoint.

> **Note**
>
> When you create a development endpoint in a virtual private cloud (VPC), AWS Glue returns only a private IP address, and the public IP address field is not populated. When you create a non-VPC development endpoint, AWS Glue returns only a public IP address.

Request

- `EndpointName` – *Required:* UTF-8 string.

  Name of the `DevEndpoint` to retrieve information for.

Response

- `DevEndpoint` – A DevEndpoint object.

  A `DevEndpoint` definition.

Errors

- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`
- `InvalidInputException`

GetDevEndpoints action (Python: get_dev_endpoints)

Retrieves all the development endpoints in this AWS account.

> **Note**
>
> When you create a development endpoint in a virtual private cloud (VPC), AWS Glue returns only a private IP address and the public IP address field is not populated. When you create a non-VPC development endpoint, AWS Glue returns only a public IP address.

Request

- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum size of information to return.
- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation call.

Response

- `DevEndpoints` – An array of DevEndpoint objects.

  A list of `DevEndpoint` definitions.
- `NextToken` – UTF-8 string.

  A continuation token, if not all `DevEndpoint` definitions have yet been returned.

Errors

- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`
- `InvalidInputException`

BatchGetDevEndpoints action (Python: batch_get_dev_endpoints)

Returns a list of resource metadata for a given list of development endpoint names. After calling the `ListDevEndpoints` operation, you can call this operation to access the data to which you have been granted permissions. This operation supports all IAM permissions, including permission conditions that uses tags.

Request

- `customerAccountId` – UTF-8 string.

  The AWS account ID.
- `DevEndpointNames` – *Required:* An array of UTF-8 strings, not less than 1 or more than 25 strings.

  The list of `DevEndpoint` names, which might be the names returned from the `ListDevEndpoint` operation.

Response

- `DevEndpoints` – An array of DevEndpoint objects.

  A list of `DevEndpoint` definitions.
- `DevEndpointsNotFound` – An array of UTF-8 strings, not less than 1 or more than 25 strings.

  A list of `DevEndpoints` not found.

Errors

- `AccessDeniedException`
- `InternalServiceException`
- `OperationTimeoutException`
- `InvalidInputException`

ListDevEndpoints action (Python: list_dev_endpoints)

Retrieves the names of all `DevEndpoint` resources in this AWS account, or the resources with the specified tag. This operation allows you to see which resources are available in your account, and their names.

This operation takes the optional `Tags` field, which you can use as a filter on the response so that tagged resources can be retrieved as a group. If you choose to use tags filtering, only resources with the tag are retrieved.

Request

- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation request.
- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum size of a list to return.
- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  Specifies to return only these tagged resources.

Response

- `DevEndpointNames` – An array of UTF-8 strings.

  The names of all the `DevEndpoints` in the account, or the `DevEndpoints` with the specified tags.
- `NextToken` – UTF-8 string.

  A continuation token, if the returned list does not contain the last metric available.

Errors

- `InvalidInputException`

- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`

Schema registry

The Schema registry API describes the data types and API related to working with schemas in AWS Glue.

Data types

- RegistryId structure
- RegistryListItem structure
- MetadataInfo structure
- OtherMetadataValueListItem structure
- SchemaListItem structure
- SchemaVersionListItem structure
- MetadataKeyValuePair structure
- SchemaVersionErrorItem structure
- ErrorDetails structure
- SchemaVersionNumber structure
- SchemaId structure

RegistryId structure

A wrapper structure that may contain the registry name and Amazon Resource Name (ARN).

Fields

- `RegistryName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #45.

  Name of the registry. Used only for lookup. One of `RegistryArn` or `RegistryName` has to be provided.

- `RegistryArn` – UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

  Arn of the registry to be updated. One of `RegistryArn` or `RegistryName` has to be provided.

RegistryListItem structure

A structure containing the details for a registry.

Fields

- `RegistryName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #45.

  The name of the registry.
- `RegistryArn` – UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

  The Amazon Resource Name (ARN) of the registry.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the registry.
- `Status` – UTF-8 string (valid values: `AVAILABLE` | `DELETING`).

  The status of the registry.
- `CreatedTime` – UTF-8 string.

  The data the registry was created.
- `UpdatedTime` – UTF-8 string.

  The date the registry was updated.

MetadataInfo structure

A structure containing metadata information for a schema version.

Fields

- `MetadataValue` – UTF-8 string, not less than 1 or more than 256 bytes long, matching the Custom string pattern #14.

  The metadata key's corresponding value.
- `CreatedTime` – UTF-8 string.

  The time at which the entry was created.
- `OtherMetadataValueList` – An array of OtherMetadataValueListItem objects.

  Other metadata belonging to the same metadata key.

OtherMetadataValueListItem structure

A structure containing other metadata for a schema version belonging to the same metadata key.

Fields

- `MetadataValue` – UTF-8 string, not less than 1 or more than 256 bytes long, matching the Custom string pattern #14.

  The metadata key's corresponding value for the other metadata belonging to the same metadata key.
- `CreatedTime` – UTF-8 string.

  The time at which the entry was created.

SchemaListItem structure

An object that contains minimal details for a schema.

Fields

- `RegistryName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #45.

  the name of the registry where the schema resides.

- `SchemaName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #45.

  The name of the schema.
- `SchemaArn` – UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

  The Amazon Resource Name (ARN) for the schema.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description for the schema.
- `SchemaStatus` – UTF-8 string (valid values: `AVAILABLE` | `PENDING` | `DELETING`).

  The status of the schema.
- `CreatedTime` – UTF-8 string.

  The date and time that a schema was created.
- `UpdatedTime` – UTF-8 string.

  The date and time that a schema was updated.

SchemaVersionListItem structure

An object containing the details about a schema version.

Fields

- `SchemaArn` – UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

  The Amazon Resource Name (ARN) of the schema.
- `SchemaVersionId` – UTF-8 string, not less than 36 or more than 36 bytes long, matching the Custom string pattern #44.

  The unique identifier of the schema version.
- `VersionNumber` – Number (long), not less than 1 or more than 100000.

  The version number of the schema.

- `Status` – UTF-8 string (valid values: `AVAILABLE` | `PENDING` | `FAILURE` | `DELETING`).

  The status of the schema version.
- `CreatedTime` – UTF-8 string.

  The date and time the schema version was created.

MetadataKeyValuePair structure

A structure containing a key value pair for metadata.

Fields

- `MetadataKey` – UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #14.

  A metadata key.
- `MetadataValue` – UTF-8 string, not less than 1 or more than 256 bytes long, matching the Custom string pattern #14.

  A metadata key's corresponding value.

SchemaVersionErrorItem structure

An object that contains the error details for an operation on a schema version.

Fields

- `VersionNumber` – Number (long), not less than 1 or more than 100000.

  The version number of the schema.
- `ErrorDetails` – An ErrorDetails object.

  The details of the error for the schema version.

ErrorDetails structure

An object containing error details.

Fields

- `ErrorCode` – UTF-8 string.

  The error code for an error.
- `ErrorMessage` – UTF-8 string.

  The error message for an error.

SchemaVersionNumber structure

A structure containing the schema version information.

Fields

- `LatestVersion` – Boolean.

  The latest version available for the schema.
- `VersionNumber` – Number (long), not less than 1 or more than 100000.

  The version number of the schema.

SchemaId structure

The unique ID of the schema in the AWS Glue schema registry.

Fields

- `SchemaArn` – UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

  The Amazon Resource Name (ARN) of the schema. One of `SchemaArn` or `SchemaName` has to be provided.
- `SchemaName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #45.

  The name of the schema. One of `SchemaArn` or `SchemaName` has to be provided.
- `RegistryName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #45.

The name of the schema registry that contains the schema.

Operations

- CreateRegistry action (Python: create_registry)
- CreateSchema action (Python: create_schema)
- GetSchema action (Python: get_schema)
- ListSchemaVersions action (Python: list_schema_versions)
- GetSchemaVersion action (Python: get_schema_version)
- GetSchemaVersionsDiff action (Python: get_schema_versions_diff)
- ListRegistries action (Python: list_registries)
- ListSchemas action (Python: list_schemas)
- RegisterSchemaVersion action (Python: register_schema_version)
- UpdateSchema action (Python: update_schema)
- CheckSchemaVersionValidity action (Python: check_schema_version_validity)
- UpdateRegistry action (Python: update_registry)
- GetSchemaByDefinition action (Python: get_schema_by_definition)
- GetRegistry action (Python: get_registry)
- PutSchemaVersionMetadata action (Python: put_schema_version_metadata)
- QuerySchemaVersionMetadata action (Python: query_schema_version_metadata)
- RemoveSchemaVersionMetadata action (Python: remove_schema_version_metadata)
- DeleteRegistry action (Python: delete_registry)
- DeleteSchema action (Python: delete_schema)
- DeleteSchemaVersions action (Python: delete_schema_versions)

CreateRegistry action (Python: create_registry)

Creates a new registry which may be used to hold a collection of schemas.

Request

- `RegistryName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #45.

  Name of the registry to be created of max length of 255, and may only contain letters, numbers, hyphen, underscore, dollar sign, or hash mark. No whitespace.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the registry. If description is not provided, there will not be any default value for this.
- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  AWS tags that contain a key value pair and may be searched by console, command line, or API.

Response

- `RegistryArn` – UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

  The Amazon Resource Name (ARN) of the newly created registry.
- `RegistryName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #45.

  The name of the registry.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the registry.
- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

The tags for the registry.

Errors

- `InvalidInputException`
- `AccessDeniedException`
- `AlreadyExistsException`
- `ResourceNumberLimitExceededException`
- `ConcurrentModificationException`
- `InternalServiceException`

CreateSchema action (Python: create_schema)

Creates a new schema set and registers the schema definition. Returns an error if the schema set already exists without actually registering the version.

When the schema set is created, a version checkpoint will be set to the first version. Compatibility mode "DISABLED" restricts any additional schema versions from being added after the first schema version. For all other compatibility modes, validation of compatibility settings will be applied only from the second version onwards when the `RegisterSchemaVersion` API is used.

When this API is called without a `RegistryId`, this will create an entry for a "default-registry" in the registry database tables, if it is not already present.

Request

- `RegistryId` – A RegistryId object.

  This is a wrapper shape to contain the registry identity fields. If this is not provided, the default registry will be used. The ARN format for the same will be: `arn:aws:glue:us-east-2:<customer id>:registry/default-registry:random-5-letter-id`.
- `SchemaName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #45.

  Name of the schema to be created of max length of 255, and may only contain letters, numbers, hyphen, underscore, dollar sign, or hash mark. No whitespace.

- `DataFormat` – *Required:* UTF-8 string (valid values: `AVRO` | `JSON` | `PROTOBUF`).

  The data format of the schema definition. Currently `AVRO`, `JSON` and `PROTOBUF` are supported.
- `Compatibility` – UTF-8 string (valid values: `NONE` | `DISABLED` | `BACKWARD` | `BACKWARD_ALL` | `FORWARD` | `FORWARD_ALL` | `FULL` | `FULL_ALL`).

  The compatibility mode of the schema. The possible values are:
  - *NONE*: No compatibility mode applies. You can use this choice in development scenarios or if you do not know the compatibility mode that you want to apply to schemas. Any new version added will be accepted without undergoing a compatibility check.
  - *DISABLED*: This compatibility choice prevents versioning for a particular schema. You can use this choice to prevent future versioning of a schema.
  - *BACKWARD*: This compatibility choice is recommended as it allows data receivers to read both the current and one previous schema version. This means that for instance, a new schema version cannot drop data fields or change the type of these fields, so they can't be read by readers using the previous version.
  - *BACKWARD_ALL*: This compatibility choice allows data receivers to read both the current and all previous schema versions. You can use this choice when you need to delete fields or add optional fields, and check compatibility against all previous schema versions.
  - *FORWARD*: This compatibility choice allows data receivers to read both the current and one next schema version, but not necessarily later versions. You can use this choice when you need to add fields or delete optional fields, but only check compatibility against the last schema version.
  - *FORWARD_ALL*: This compatibility choice allows data receivers to read written by producers of any new registered schema. You can use this choice when you need to add fields or delete optional fields, and check compatibility against all previous schema versions.
  - *FULL*: This compatibility choice allows data receivers to read data written by producers using the previous or next version of the schema, but not necessarily earlier or later versions. You can use this choice when you need to add or remove optional fields, but only check compatibility against the last schema version.
  - *FULL_ALL*: This compatibility choice allows data receivers to read data written by producers using all previous schema versions. You can use this choice when you need to add or remove optional fields, and check compatibility against all previous schema versions.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

An optional description of the schema. If description is not provided, there will not be any automatic default value for this.

- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  AWS tags that contain a key value pair and may be searched by console, command line, or API. If specified, follows the AWS tags-on-create pattern.
- `SchemaDefinition` – UTF-8 string, not less than 1 or more than 170000 bytes long, matching the Custom string pattern #13.

  The schema definition using the `DataFormat` setting for `SchemaName`.

Response

- `RegistryName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #45.

  The name of the registry.
- `RegistryArn` – UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

  The Amazon Resource Name (ARN) of the registry.
- `SchemaName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #45.

  The name of the schema.
- `SchemaArn` – UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

  The Amazon Resource Name (ARN) of the schema.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the schema if specified when created.

- `DataFormat` – UTF-8 string (valid values: `AVRO` | `JSON` | `PROTOBUF`).

  The data format of the schema definition. Currently `AVRO`, `JSON` and `PROTOBUF` are supported.

- `Compatibility` – UTF-8 string (valid values: `NONE` | `DISABLED` | `BACKWARD` | `BACKWARD_ALL` | `FORWARD` | `FORWARD_ALL` | `FULL` | `FULL_ALL`).

  The schema compatibility mode.

- `SchemaCheckpoint` – Number (long), not less than 1 or more than 100000.

  The version number of the checkpoint (the last time the compatibility mode was changed).

- `LatestSchemaVersion` – Number (long), not less than 1 or more than 100000.

  The latest version of the schema associated with the returned schema definition.

- `NextSchemaVersion` – Number (long), not less than 1 or more than 100000.

  The next version of the schema associated with the returned schema definition.

- `SchemaStatus` – UTF-8 string (valid values: `AVAILABLE` | `PENDING` | `DELETING`).

  The status of the schema.

- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  The tags for the schema.

- `SchemaVersionId` – UTF-8 string, not less than 36 or more than 36 bytes long, matching the Custom string pattern #44.

  The unique identifier of the first schema version.

- `SchemaVersionStatus` – UTF-8 string (valid values: `AVAILABLE` | `PENDING` | `FAILURE` | `DELETING`).

  The status of the first schema version created.

Errors

- `InvalidInputException`

- `AccessDeniedException`
- `EntityNotFoundException`
- `AlreadyExistsException`
- `ResourceNumberLimitExceededException`
- `ConcurrentModificationException`
- `InternalServiceException`

GetSchema action (Python: get_schema)

Describes the specified schema in detail.

Request

- `SchemaId` – *Required:* A SchemaId object.

  This is a wrapper structure to contain schema identity fields. The structure contains:

  - SchemaId$SchemaArn: The Amazon Resource Name (ARN) of the schema. Either `SchemaArn` or `SchemaName` and `RegistryName` has to be provided.
  - SchemaId$SchemaName: The name of the schema. Either `SchemaArn` or `SchemaName` and `RegistryName` has to be provided.

Response

- `RegistryName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #45.

  The name of the registry.
- `RegistryArn` – UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

  The Amazon Resource Name (ARN) of the registry.
- `SchemaName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #45.

  The name of the schema.

- `SchemaArn` – UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

  The Amazon Resource Name (ARN) of the schema.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of schema if specified when created
- `DataFormat` – UTF-8 string (valid values: `AVRO` | `JSON` | `PROTOBUF`).

  The data format of the schema definition. Currently `AVRO`, `JSON` and `PROTOBUF` are supported.
- `Compatibility` – UTF-8 string (valid values: `NONE` | `DISABLED` | `BACKWARD` | `BACKWARD_ALL` | `FORWARD` | `FORWARD_ALL` | `FULL` | `FULL_ALL`).

  The compatibility mode of the schema.
- `SchemaCheckpoint` – Number (long), not less than 1 or more than 100000.

  The version number of the checkpoint (the last time the compatibility mode was changed).
- `LatestSchemaVersion` – Number (long), not less than 1 or more than 100000.

  The latest version of the schema associated with the returned schema definition.
- `NextSchemaVersion` – Number (long), not less than 1 or more than 100000.

  The next version of the schema associated with the returned schema definition.
- `SchemaStatus` – UTF-8 string (valid values: `AVAILABLE` | `PENDING` | `DELETING`).

  The status of the schema.
- `CreatedTime` – UTF-8 string.

  The date and time the schema was created.
- `UpdatedTime` – UTF-8 string.

  The date and time the schema was updated.

Errors

- `InvalidInputException`
- `AccessDeniedException`

- `EntityNotFoundException`
- `InternalServiceException`

ListSchemaVersions action (Python: list_schema_versions)

Returns a list of schema versions that you have created, with minimal information. Schema versions in Deleted status will not be included in the results. Empty results will be returned if there are no schema versions available.

Request

- `SchemaId` – *Required:* A SchemaId object.

  This is a wrapper structure to contain schema identity fields. The structure contains:

  - SchemaId$SchemaArn: The Amazon Resource Name (ARN) of the schema. Either `SchemaArn` or `SchemaName` and `RegistryName` has to be provided.
  - SchemaId$SchemaName: The name of the schema. Either `SchemaArn` or `SchemaName` and `RegistryName` has to be provided.
- `MaxResults` – Number (integer), not less than 1 or more than 100.

  Maximum number of results required per page. If the value is not supplied, this will be defaulted to 25 per page.
- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation call.

Response

- `Schemas` – An array of SchemaVersionListItem objects.

  An array of `SchemaVersionList` objects containing details of each schema version.
- `NextToken` – UTF-8 string.

  A continuation token for paginating the returned list of tokens, returned if the current segment of the list is not the last.

Errors

- `InvalidInputException`
- `AccessDeniedException`
- `EntityNotFoundException`
- `InternalServiceException`

GetSchemaVersion action (Python: get_schema_version)

Get the specified schema by its unique ID assigned when a version of the schema is created or registered. Schema versions in Deleted status will not be included in the results.

Request

- `SchemaId` – A SchemaId object.

  This is a wrapper structure to contain schema identity fields. The structure contains:

  - SchemaId$SchemaArn: The Amazon Resource Name (ARN) of the schema. Either `SchemaArn` or `SchemaName` and `RegistryName` has to be provided.
  - SchemaId$SchemaName: The name of the schema. Either `SchemaArn` or `SchemaName` and `RegistryName` has to be provided.
- `SchemaVersionId` – UTF-8 string, not less than 36 or more than 36 bytes long, matching the Custom string pattern #44.

  The `SchemaVersionId` of the schema version. This field is required for fetching by schema ID. Either this or the `SchemaId` wrapper has to be provided.
- `SchemaVersionNumber` – A SchemaVersionNumber object.

  The version number of the schema.

Response

- `SchemaVersionId` – UTF-8 string, not less than 36 or more than 36 bytes long, matching the Custom string pattern #44.

  The `SchemaVersionId` of the schema version.

- `SchemaDefinition` – UTF-8 string, not less than 1 or more than 170000 bytes long, matching the Custom string pattern #13.

  The schema definition for the schema ID.
- `DataFormat` – UTF-8 string (valid values: `AVRO` | `JSON` | `PROTOBUF`).

  The data format of the schema definition. Currently `AVRO`, `JSON` and `PROTOBUF` are supported.
- `SchemaArn` – UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

  The Amazon Resource Name (ARN) of the schema.
- `VersionNumber` – Number (long), not less than 1 or more than 100000.

  The version number of the schema.
- `Status` – UTF-8 string (valid values: `AVAILABLE` | `PENDING` | `FAILURE` | `DELETING`).

  The status of the schema version.
- `CreatedTime` – UTF-8 string.

  The date and time the schema version was created.

Errors

- `InvalidInputException`
- `AccessDeniedException`
- `EntityNotFoundException`
- `InternalServiceException`

GetSchemaVersionsDiff action (Python: get_schema_versions_diff)

Fetches the schema version difference in the specified difference type between two stored schema versions in the Schema Registry.

This API allows you to compare two schema versions between two schema definitions under the same schema.

Request

- `SchemaId` – *Required:* A SchemaId object.

  This is a wrapper structure to contain schema identity fields. The structure contains:

  - SchemaId$SchemaArn: The Amazon Resource Name (ARN) of the schema. One of `SchemaArn` or `SchemaName` has to be provided.
  - SchemaId$SchemaName: The name of the schema. One of `SchemaArn` or `SchemaName` has to be provided.
- `FirstSchemaVersionNumber` – *Required:* A SchemaVersionNumber object.

  The first of the two schema versions to be compared.
- `SecondSchemaVersionNumber` – *Required:* A SchemaVersionNumber object.

  The second of the two schema versions to be compared.
- `SchemaDiffType` – *Required:* UTF-8 string (valid values: `SYNTAX_DIFF`).

  Refers to `SYNTAX_DIFF`, which is the currently supported diff type.

Response

- `Diff` – UTF-8 string, not less than 1 or more than 340000 bytes long, matching the Custom string pattern #13.

  The difference between schemas as a string in JsonPatch format.

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `AccessDeniedException`
- `InternalServiceException`

ListRegistries action (Python: list_registries)

Returns a list of registries that you have created, with minimal registry information. Registries in the `Deleting` status will not be included in the results. Empty results will be returned if there are no registries available.

Request

- `MaxResults` – Number (integer), not less than 1 or more than 100.

  Maximum number of results required per page. If the value is not supplied, this will be defaulted to 25 per page.
- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation call.

Response

- `Registries` – An array of RegistryListItem objects.

  An array of `RegistryDetailedListItem` objects containing minimal details of each registry.
- `NextToken` – UTF-8 string.

  A continuation token for paginating the returned list of tokens, returned if the current segment of the list is not the last.

Errors

- `InvalidInputException`
- `AccessDeniedException`
- `InternalServiceException`

ListSchemas action (Python: list_schemas)

Returns a list of schemas with minimal details. Schemas in Deleting status will not be included in the results. Empty results will be returned if there are no schemas available.

When the `RegistryId` is not provided, all the schemas across registries will be part of the API response.

Request

- `RegistryId` – A RegistryId object.

  A wrapper structure that may contain the registry name and Amazon Resource Name (ARN).
- `MaxResults` – Number (integer), not less than 1 or more than 100.

  Maximum number of results required per page. If the value is not supplied, this will be defaulted to 25 per page.
- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation call.

Response

- `Schemas` – An array of SchemaListItem objects.

  An array of `SchemaListItem` objects containing details of each schema.
- `NextToken` – UTF-8 string.

  A continuation token for paginating the returned list of tokens, returned if the current segment of the list is not the last.

Errors

- `InvalidInputException`
- `AccessDeniedException`
- `EntityNotFoundException`
- `InternalServiceException`

RegisterSchemaVersion action (Python: register_schema_version)

Adds a new version to the existing schema. Returns an error if new version of schema does not meet the compatibility requirements of the schema set. This API will not create a new schema set and will return a 404 error if the schema set is not already present in the Schema Registry.

If this is the first schema definition to be registered in the Schema Registry, this API will store the schema version and return immediately. Otherwise, this call has the potential to run longer than other operations due to compatibility modes. You can call the `GetSchemaVersion` API with the `SchemaVersionId` to check compatibility modes.

If the same schema definition is already stored in Schema Registry as a version, the schema ID of the existing schema is returned to the caller.

Request

- `SchemaId` – *Required:* A SchemaId object.

  This is a wrapper structure to contain schema identity fields. The structure contains:
  - SchemaId$SchemaArn: The Amazon Resource Name (ARN) of the schema. Either `SchemaArn` or `SchemaName` and `RegistryName` has to be provided.
  - SchemaId$SchemaName: The name of the schema. Either `SchemaArn` or `SchemaName` and `RegistryName` has to be provided.
- `SchemaDefinition` – *Required:* UTF-8 string, not less than 1 or more than 170000 bytes long, matching the Custom string pattern #13.

  The schema definition using the `DataFormat` setting for the `SchemaName`.

Response

- `SchemaVersionId` – UTF-8 string, not less than 36 or more than 36 bytes long, matching the Custom string pattern #44.

  The unique ID that represents the version of this schema.
- `VersionNumber` – Number (long), not less than 1 or more than 100000.

  The version of this schema (for sync flow only, in case this is the first version).
- `Status` – UTF-8 string (valid values: `AVAILABLE` | `PENDING` | `FAILURE` | `DELETING`).

The status of the schema version.

Errors

- `InvalidInputException`
- `AccessDeniedException`
- `EntityNotFoundException`
- `ResourceNumberLimitExceededException`
- `ConcurrentModificationException`
- `InternalServiceException`

UpdateSchema action (Python: update_schema)

Updates the description, compatibility setting, or version checkpoint for a schema set.

For updating the compatibility setting, the call will not validate compatibility for the entire set of schema versions with the new compatibility setting. If the value for `Compatibility` is provided, the `VersionNumber` (a checkpoint) is also required. The API will validate the checkpoint version number for consistency.

If the value for the `VersionNumber` (checkpoint) is provided, `Compatibility` is optional and this can be used to set/reset a checkpoint for the schema.

This update will happen only if the schema is in the AVAILABLE state.

Request

- `SchemaId` – *Required:* A SchemaId object.

  This is a wrapper structure to contain schema identity fields. The structure contains:

  - SchemaId$SchemaArn: The Amazon Resource Name (ARN) of the schema. One of `SchemaArn` or `SchemaName` has to be provided.
  - SchemaId$SchemaName: The name of the schema. One of `SchemaArn` or `SchemaName` has to be provided.
- `SchemaVersionNumber` – A SchemaVersionNumber object.

Version number required for check pointing. One of `VersionNumber` or `Compatibility` has to be provided.

- `Compatibility` – UTF-8 string (valid values: `NONE` | `DISABLED` | `BACKWARD` | `BACKWARD_ALL` | `FORWARD` | `FORWARD_ALL` | `FULL` | `FULL_ALL`).

  The new compatibility setting for the schema.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  The new description for the schema.

Response

- `SchemaArn` – UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

  The Amazon Resource Name (ARN) of the schema.
- `SchemaName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #45.

  The name of the schema.
- `RegistryName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #45.

  The name of the registry that contains the schema.

Errors

- `InvalidInputException`
- `AccessDeniedException`
- `EntityNotFoundException`
- `ConcurrentModificationException`
- `InternalServiceException`

CheckSchemaVersionValidity action (Python: check_schema_version_validity)

Validates the supplied schema. This call has no side effects, it simply validates using the supplied schema using `DataFormat` as the format. Since it does not take a schema set name, no compatibility checks are performed.

Request

- `DataFormat` – *Required:* UTF-8 string (valid values: `AVRO` | `JSON` | `PROTOBUF`).

  The data format of the schema definition. Currently `AVRO`, `JSON` and `PROTOBUF` are supported.
- `SchemaDefinition` – *Required:* UTF-8 string, not less than 1 or more than 170000 bytes long, matching the Custom string pattern #13.

  The definition of the schema that has to be validated.

Response

- `Valid` – Boolean.

  Return true, if the schema is valid and false otherwise.
- `Error` – UTF-8 string, not less than 1 or more than 5000 bytes long.

  A validation failure error message.

Errors

- `InvalidInputException`
- `AccessDeniedException`
- `InternalServiceException`

UpdateRegistry action (Python: update_registry)

Updates an existing registry which is used to hold a collection of schemas. The updated properties relate to the registry, and do not modify any of the schemas within the registry.

Request

- `RegistryId` – *Required:* A RegistryId object.

  This is a wrapper structure that may contain the registry name and Amazon Resource Name (ARN).
- `Description` – *Required:* Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the registry. If description is not provided, this field will not be updated.

Response

- `RegistryName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #45.

  The name of the updated registry.
- `RegistryArn` – UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

  The Amazon Resource name (ARN) of the updated registry.

Errors

- `InvalidInputException`
- `AccessDeniedException`
- `EntityNotFoundException`
- `ConcurrentModificationException`
- `InternalServiceException`

GetSchemaByDefinition action (Python: get_schema_by_definition)

Retrieves a schema by the `SchemaDefinition`. The schema definition is sent to the Schema Registry, canonicalized, and hashed. If the hash is matched within the scope of the `SchemaName` or ARN (or the default registry, if none is supplied), that schema's metadata is returned. Otherwise, a 404 or NotFound error is returned. Schema versions in `Deleted` statuses will not be included in the results.

Request

- `SchemaId` – *Required:* A SchemaId object.

  This is a wrapper structure to contain schema identity fields. The structure contains:

  - SchemaId$SchemaArn: The Amazon Resource Name (ARN) of the schema. One of `SchemaArn` or `SchemaName` has to be provided.
  - SchemaId$SchemaName: The name of the schema. One of `SchemaArn` or `SchemaName` has to be provided.
- `SchemaDefinition` – *Required:* UTF-8 string, not less than 1 or more than 170000 bytes long, matching the Custom string pattern #13.

  The definition of the schema for which schema details are required.

Response

- `SchemaVersionId` – UTF-8 string, not less than 36 or more than 36 bytes long, matching the Custom string pattern #44.

  The schema ID of the schema version.
- `SchemaArn` – UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

  The Amazon Resource Name (ARN) of the schema.
- `DataFormat` – UTF-8 string (valid values: `AVRO` | `JSON` | `PROTOBUF`).

  The data format of the schema definition. Currently `AVRO`, `JSON` and `PROTOBUF` are supported.
- `Status` – UTF-8 string (valid values: `AVAILABLE` | `PENDING` | `FAILURE` | `DELETING`).

  The status of the schema version.
- `CreatedTime` – UTF-8 string.

  The date and time the schema was created.

Errors

- `InvalidInputException`

- `AccessDeniedException`
- `EntityNotFoundException`
- `InternalServiceException`

GetRegistry action (Python: get_registry)

Describes the specified registry in detail.

Request

- `RegistryId` – *Required:* A RegistryId object.

  This is a wrapper structure that may contain the registry name and Amazon Resource Name (ARN).

Response

- `RegistryName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #45.

  The name of the registry.
- `RegistryArn` – UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

  The Amazon Resource Name (ARN) of the registry.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the registry.
- `Status` – UTF-8 string (valid values: `AVAILABLE` | `DELETING`).

  The status of the registry.
- `CreatedTime` – UTF-8 string.

  The date and time the registry was created.
- `UpdatedTime` – UTF-8 string.

  The date and time the registry was updated.

Errors

- `InvalidInputException`
- `AccessDeniedException`
- `EntityNotFoundException`
- `InternalServiceException`

PutSchemaVersionMetadata action (Python: put_schema_version_metadata)

Puts the metadata key value pair for a specified schema version ID. A maximum of 10 key value pairs will be allowed per schema version. They can be added over one or more calls.

Request

- `SchemaId` – A SchemaId object.

  The unique ID for the schema.
- `SchemaVersionNumber` – A SchemaVersionNumber object.

  The version number of the schema.
- `SchemaVersionId` – UTF-8 string, not less than 36 or more than 36 bytes long, matching the Custom string pattern #44.

  The unique version ID of the schema version.
- `MetadataKeyValue` – *Required:* A MetadataKeyValuePair object.

  The metadata key's corresponding value.

Response

- `SchemaArn` – UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

  The Amazon Resource Name (ARN) for the schema.
- `SchemaName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #45.

The name for the schema.

- `RegistryName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #45.

  The name for the registry.
- `LatestVersion` – Boolean.

  The latest version of the schema.
- `VersionNumber` – Number (long), not less than 1 or more than 100000.

  The version number of the schema.
- `SchemaVersionId` – UTF-8 string, not less than 36 or more than 36 bytes long, matching the Custom string pattern #44.

  The unique version ID of the schema version.
- `MetadataKey` – UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #14.

  The metadata key.
- `MetadataValue` – UTF-8 string, not less than 1 or more than 256 bytes long, matching the Custom string pattern #14.

  The value of the metadata key.

Errors

- `InvalidInputException`
- `AccessDeniedException`
- `AlreadyExistsException`
- `EntityNotFoundException`
- `ResourceNumberLimitExceededException`

QuerySchemaVersionMetadata action (Python: query_schema_version_metadata)

Queries for the schema version metadata information.

Request

- `SchemaId` – A SchemaId object.

  A wrapper structure that may contain the schema name and Amazon Resource Name (ARN).
- `SchemaVersionNumber` – A SchemaVersionNumber object.

  The version number of the schema.
- `SchemaVersionId` – UTF-8 string, not less than 36 or more than 36 bytes long, matching the Custom string pattern #44.

  The unique version ID of the schema version.
- `MetadataList` – An array of MetadataKeyValuePair objects.

  Search key-value pairs for metadata, if they are not provided all the metadata information will be fetched.
- `MaxResults` – Number (integer), not less than 1 or more than 50.

  Maximum number of results required per page. If the value is not supplied, this will be defaulted to 25 per page.
- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation call.

Response

- `MetadataInfoMap` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #14.

  Each value is a A MetadataInfo object.

  A map of a metadata key and associated values.

- `SchemaVersionId` – UTF-8 string, not less than 36 or more than 36 bytes long, matching the Custom string pattern #44.

  The unique version ID of the schema version.
- `NextToken` – UTF-8 string.

  A continuation token for paginating the returned list of tokens, returned if the current segment of the list is not the last.

Errors

- `InvalidInputException`
- `AccessDeniedException`
- `EntityNotFoundException`

RemoveSchemaVersionMetadata action (Python: remove_schema_version_metadata)

Removes a key value pair from the schema version metadata for the specified schema version ID.

Request

- `SchemaId` – A SchemaId object.

  A wrapper structure that may contain the schema name and Amazon Resource Name (ARN).
- `SchemaVersionNumber` – A SchemaVersionNumber object.

  The version number of the schema.
- `SchemaVersionId` – UTF-8 string, not less than 36 or more than 36 bytes long, matching the Custom string pattern #44.

  The unique version ID of the schema version.
- `MetadataKeyValue` – *Required:* A MetadataKeyValuePair object.

  The value of the metadata key.

Response

- `SchemaArn` – UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

  The Amazon Resource Name (ARN) of the schema.
- `SchemaName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #45.

  The name of the schema.
- `RegistryName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #45.

  The name of the registry.
- `LatestVersion` – Boolean.

  The latest version of the schema.
- `VersionNumber` – Number (long), not less than 1 or more than 100000.

  The version number of the schema.
- `SchemaVersionId` – UTF-8 string, not less than 36 or more than 36 bytes long, matching the Custom string pattern #44.

  The version ID for the schema version.
- `MetadataKey` – UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #14.

  The metadata key.
- `MetadataValue` – UTF-8 string, not less than 1 or more than 256 bytes long, matching the Custom string pattern #14.

  The value of the metadata key.

Errors

- `InvalidInputException`
- `AccessDeniedException`
- `EntityNotFoundException`

DeleteRegistry action (Python: delete_registry)

Delete the entire registry including schema and all of its versions. To get the status of the delete operation, you can call the `GetRegistry` API after the asynchronous call. Deleting a registry will deactivate all online operations for the registry such as the `UpdateRegistry`, `CreateSchema`, `UpdateSchema`, and `RegisterSchemaVersion` APIs.

Request

- `RegistryId` – *Required:* A RegistryId object.

  This is a wrapper structure that may contain the registry name and Amazon Resource Name (ARN).

Response

- `RegistryName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #45.

  The name of the registry being deleted.
- `RegistryArn` – UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

  The Amazon Resource Name (ARN) of the registry being deleted.
- `Status` – UTF-8 string (valid values: `AVAILABLE` | `DELETING`).

  The status of the registry. A successful operation will return the `Deleting` status.

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `AccessDeniedException`
- `ConcurrentModificationException`

DeleteSchema action (Python: delete_schema)

Deletes the entire schema set, including the schema set and all of its versions. To get the status of the delete operation, you can call `GetSchema` API after the asynchronous call. Deleting a registry will deactivate all online operations for the schema, such as the `GetSchemaByDefinition`, and `RegisterSchemaVersion` APIs.

Request

- `SchemaId` – *Required:* A SchemaId object.

  This is a wrapper structure that may contain the schema name and Amazon Resource Name (ARN).

Response

- `SchemaArn` – UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

  The Amazon Resource Name (ARN) of the schema being deleted.
- `SchemaName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #45.

  The name of the schema being deleted.
- `Status` – UTF-8 string (valid values: `AVAILABLE` | `PENDING` | `DELETING`).

  The status of the schema.

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `AccessDeniedException`
- `ConcurrentModificationException`

DeleteSchemaVersions action (Python: delete_schema_versions)

Remove versions from the specified schema. A version number or range may be supplied. If the compatibility mode forbids deleting of a version that is necessary, such as BACKWARDS_FULL, an error is returned. Calling the `GetSchemaVersions` API after this call will list the status of the deleted versions.

When the range of version numbers contain check pointed version, the API will return a 409 conflict and will not proceed with the deletion. You have to remove the checkpoint first using the `DeleteSchemaCheckpoint` API before using this API.

You cannot use the `DeleteSchemaVersions` API to delete the first schema version in the schema set. The first schema version can only be deleted by the `DeleteSchema` API. This operation will also delete the attached `SchemaVersionMetadata` under the schema versions. Hard deletes will be enforced on the database.

If the compatibility mode forbids deleting of a version that is necessary, such as BACKWARDS_FULL, an error is returned.

Request

- `SchemaId` – *Required:* A SchemaId object.

  This is a wrapper structure that may contain the schema name and Amazon Resource Name (ARN).

- `Versions` – *Required:* UTF-8 string, not less than 1 or more than 100000 bytes long, matching the Custom string pattern #15.

  A version range may be supplied which may be of the format:

  - a single version number, 5
  - a range, 5-8 : deletes versions 5, 6, 7, 8

Response

- `SchemaVersionErrors` – An array of SchemaVersionErrorItem objects.

  A list of `SchemaVersionErrorItem` objects, each containing an error and schema version.

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `AccessDeniedException`
- `ConcurrentModificationException`

Workflows

The Workflows API describes the data types and API related to creating, updating, or viewing workflows in AWS Glue. Job run history is accessible for 90 days for your workflow and job run.

Data types

- JobNodeDetails structure
- CrawlerNodeDetails structure
- TriggerNodeDetails structure
- Crawl structure
- Node structure
- Edge structure
- Workflow structure
- WorkflowGraph structure
- WorkflowRun structure
- WorkflowRunStatistics structure
- StartingEventBatchCondition structure
- Blueprint structure
- BlueprintDetails structure
- LastActiveDefinition structure
- BlueprintRun structure

JobNodeDetails structure

The details of a Job node present in the workflow.

Fields

- `JobRuns` – An array of JobRun objects.

  The information for the job runs represented by the job node.

CrawlerNodeDetails structure

The details of a Crawler node present in the workflow.

Fields

- `Crawls` – An array of Crawl objects.

  A list of crawls represented by the crawl node.

TriggerNodeDetails structure

The details of a Trigger node present in the workflow.

Fields

- `Trigger` – A Trigger object.

  The information of the trigger represented by the trigger node.

Crawl structure

The details of a crawl in the workflow.

Fields

- `State` – UTF-8 string (valid values: `RUNNING` | `CANCELLING` | `CANCELLED` | `SUCCEEDED` | `FAILED` | `ERROR`).

  The state of the crawler.
- `StartedOn` – Timestamp.

  The date and time on which the crawl started.
- `CompletedOn` – Timestamp.

The date and time on which the crawl completed.

- `ErrorMessage` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  The error message associated with the crawl.

- `LogGroup` – UTF-8 string, not less than 1 or more than 512 bytes long, matching the Log group string pattern.

  The log group associated with the crawl.

- `LogStream` – UTF-8 string, not less than 1 or more than 512 bytes long, matching the Log-stream string pattern.

  The log stream associated with the crawl.

Node structure

A node represents an AWS Glue component (trigger, crawler, or job) on a workflow graph.

Fields

- `Type` – UTF-8 string (valid values: `CRAWLER` | `JOB` | `TRIGGER`).

  The type of AWS Glue component represented by the node.

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the AWS Glue component represented by the node.

- `UniqueId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique Id assigned to the node within the workflow.

- `TriggerDetails` – A TriggerNodeDetails object.

  Details of the Trigger when the node represents a Trigger.

- `JobDetails` – A JobNodeDetails object.

  Details of the Job when the node represents a Job.

- `CrawlerDetails` – A CrawlerNodeDetails object.

  Details of the crawler when the node represents a crawler.

Edge structure

An edge represents a directed connection between two AWS Glue components that are part of the workflow the edge belongs to.

Fields

- `SourceId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique of the node within the workflow where the edge starts.
- `DestinationId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique of the node within the workflow where the edge ends.

Workflow structure

A workflow is a collection of multiple dependent AWS Glue jobs and crawlers that are run to complete a complex ETL task. A workflow manages the execution and monitoring of all its jobs and crawlers.

Fields

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the workflow.
- `Description` – UTF-8 string.

  A description of the workflow.
- `DefaultRunProperties` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

Each value is a UTF-8 string.

A collection of properties to be used as part of each execution of the workflow. The run properties are made available to each job in the workflow. A job can modify the properties for the next jobs in the flow.

- `CreatedOn` – Timestamp.

  The date and time when the workflow was created.
- `LastModifiedOn` – Timestamp.

  The date and time when the workflow was last modified.
- `LastRun` – A WorkflowRun object.

  The information about the last execution of the workflow.
- `Graph` – A WorkflowGraph object.

  The graph representing all the AWS Glue components that belong to the workflow as nodes and directed connections between them as edges.
- `CreationStatus` – UTF-8 string (valid values: `CREATING` | `CREATED` | `CREATION_FAILED`).

  The creation status of the workflow.
- `MaxConcurrentRuns` – Number (integer).

  You can use this parameter to prevent unwanted multiple updates to data, to control costs, or in some cases, to prevent exceeding the maximum number of concurrent runs of any of the component jobs. If you leave this parameter blank, there is no limit to the number of concurrent workflow runs.
- `BlueprintDetails` – A BlueprintDetails object.

  This structure indicates the details of the blueprint that this particular workflow is created from.

WorkflowGraph structure

A workflow graph represents the complete workflow containing all the AWS Glue components present in the workflow and all the directed connections between them.

Fields

- `Nodes` – An array of Node objects.

  A list of the the AWS Glue components belong to the workflow represented as nodes.
- `Edges` – An array of Edge objects.

  A list of all the directed connections between the nodes belonging to the workflow.

WorkflowRun structure

A workflow run is an execution of a workflow providing all the runtime information.

Fields

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Name of the workflow that was run.
- `WorkflowRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of this workflow run.
- `PreviousRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the previous workflow run.
- `WorkflowRunProperties` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a UTF-8 string.

  The workflow run properties which were set during the run.
- `StartedOn` – Timestamp.

  The date and time when the workflow run was started.
- `CompletedOn` – Timestamp.

The date and time when the workflow run completed.

- `Status` – UTF-8 string (valid values: `RUNNING` | `COMPLETED` | `STOPPING` | `STOPPED` | `ERROR`).

  The status of the workflow run.
- `ErrorMessage` – UTF-8 string.

  This error message describes any error that may have occurred in starting the workflow run. Currently the only error message is "Concurrent runs exceeded for workflow: `foo`."
- `Statistics` – A [WorkflowRunStatistics]() object.

  The statistics of the run.
- `Graph` – A [WorkflowGraph]() object.

  The graph representing all the AWS Glue components that belong to the workflow as nodes and directed connections between them as edges.
- `StartingEventBatchCondition` – A [StartingEventBatchCondition]() object.

  The batch condition that started the workflow run.

WorkflowRunStatistics structure

Workflow run statistics provides statistics about the workflow run.

Fields

- `TotalActions` – Number (integer).

  Total number of Actions in the workflow run.
- `TimeoutActions` – Number (integer).

  Total number of Actions that timed out.
- `FailedActions` – Number (integer).

  Total number of Actions that have failed.
- `StoppedActions` – Number (integer).

  Total number of Actions that have stopped.

- `SucceededActions` – Number (integer).

  Total number of Actions that have succeeded.
- `RunningActions` – Number (integer).

  Total number Actions in running state.
- `ErroredActions` – Number (integer).

  Indicates the count of job runs in the ERROR state in the workflow run.
- `WaitingActions` – Number (integer).

  Indicates the count of job runs in WAITING state in the workflow run.

StartingEventBatchCondition structure

The batch condition that started the workflow run. Either the number of events in the batch size arrived, in which case the BatchSize member is non-zero, or the batch window expired, in which case the BatchWindow member is non-zero.

Fields

- `BatchSize` – Number (integer).

  Number of events in the batch.
- `BatchWindow` – Number (integer).

  Duration of the batch window in seconds.

Blueprint structure

The details of a blueprint.

Fields

- `Name` – UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #31.

  The name of the blueprint.
- `Description` – UTF-8 string, not less than 1 or more than 512 bytes long.

The description of the blueprint.

- `CreatedOn` – Timestamp.

  The date and time the blueprint was registered.
- `LastModifiedOn` – Timestamp.

  The date and time the blueprint was last modified.
- `ParameterSpec` – UTF-8 string, not less than 1 or more than 131072 bytes long.

  A JSON string that indicates the list of parameter specifications for the blueprint.
- `BlueprintLocation` – UTF-8 string.

  Specifies the path in Amazon S3 where the blueprint is published.
- `BlueprintServiceLocation` – UTF-8 string.

  Specifies a path in Amazon S3 where the blueprint is copied when you call `CreateBlueprint/UpdateBlueprint` to register the blueprint in AWS Glue.
- `Status` – UTF-8 string (valid values: `CREATING` | `ACTIVE` | `UPDATING` | `FAILED`).

  The status of the blueprint registration.
  - Creating — The blueprint registration is in progress.
  - Active — The blueprint has been successfully registered.
  - Updating — An update to the blueprint registration is in progress.
  - Failed — The blueprint registration failed.
- `ErrorMessage` – UTF-8 string.

  An error message.
- `LastActiveDefinition` – A LastActiveDefinition object.

  When there are multiple versions of a blueprint and the latest version has some errors, this attribute indicates the last successful blueprint definition that is available with the service.

BlueprintDetails structure

The details of a blueprint.

Fields

- `BlueprintName` – UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #31.

  The name of the blueprint.
- `RunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The run ID for this blueprint.

LastActiveDefinition structure

When there are multiple versions of a blueprint and the latest version has some errors, this attribute indicates the last successful blueprint definition that is available with the service.

Fields

- `Description` – UTF-8 string, not less than 1 or more than 512 bytes long.

  The description of the blueprint.
- `LastModifiedOn` – Timestamp.

  The date and time the blueprint was last modified.
- `ParameterSpec` – UTF-8 string, not less than 1 or more than 131072 bytes long.

  A JSON string specifying the parameters for the blueprint.
- `BlueprintLocation` – UTF-8 string.

  Specifies a path in Amazon S3 where the blueprint is published by the AWS Glue developer.
- `BlueprintServiceLocation` – UTF-8 string.

  Specifies a path in Amazon S3 where the blueprint is copied when you create or update the blueprint.

BlueprintRun structure

The details of a blueprint run.

Fields

- `BlueprintName` – UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #31.

  The name of the blueprint.
- `RunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The run ID for this blueprint run.
- `WorkflowName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of a workflow that is created as a result of a successful blueprint run. If a blueprint run has an error, there will not be a workflow created.
- `State` – UTF-8 string (valid values: `RUNNING` | `SUCCEEDED` | `FAILED` | `ROLLING_BACK`).

  The state of the blueprint run. Possible values are:
  - Running — The blueprint run is in progress.
  - Succeeded — The blueprint run completed successfully.
  - Failed — The blueprint run failed and rollback is complete.
  - Rolling Back — The blueprint run failed and rollback is in progress.
- `StartedOn` – Timestamp.

  The date and time that the blueprint run started.
- `CompletedOn` – Timestamp.

  The date and time that the blueprint run completed.
- `ErrorMessage` – UTF-8 string.

  Indicates any errors that are seen while running the blueprint.
- `RollbackErrorMessage` – UTF-8 string.

  If there are any errors while creating the entities of a workflow, we try to roll back the created entities until that point and delete them. This attribute indicates the errors seen while trying to delete the entities that are created.
- `Parameters` – UTF-8 string, not less than 1 or more than 131072 bytes long.

The blueprint parameters as a string. You will have to provide a value for each key that is required from the parameter spec that is defined in the `Blueprint$ParameterSpec`.

- `RoleArn` – UTF-8 string, not less than 1 or more than 1024 bytes long, matching the Custom string pattern #30.

  The role ARN. This role will be assumed by the AWS Glue service and will be used to create the workflow and other entities of a workflow.

Operations

- CreateWorkflow action (Python: create_workflow)
- UpdateWorkflow action (Python: update_workflow)
- DeleteWorkflow action (Python: delete_workflow)
- GetWorkflow action (Python: get_workflow)
- ListWorkflows action (Python: list_workflows)
- BatchGetWorkflows action (Python: batch_get_workflows)
- GetWorkflowRun action (Python: get_workflow_run)
- GetWorkflowRuns action (Python: get_workflow_runs)
- GetWorkflowRunProperties action (Python: get_workflow_run_properties)
- PutWorkflowRunProperties action (Python: put_workflow_run_properties)
- CreateBlueprint action (Python: create_blueprint)
- UpdateBlueprint action (Python: update_blueprint)
- DeleteBlueprint action (Python: delete_blueprint)
- ListBlueprints action (Python: list_blueprints)
- BatchGetBlueprints action (Python: batch_get_blueprints)
- StartBlueprintRun action (Python: start_blueprint_run)
- GetBlueprintRun action (Python: get_blueprint_run)
- GetBlueprintRuns action (Python: get_blueprint_runs)
- StartWorkflowRun action (Python: start_workflow_run)
- StopWorkflowRun action (Python: stop_workflow_run)

- ResumeWorkflowRun action (Python: resume_workflow_run)

CreateWorkflow action (Python: create_workflow)

Creates a new workflow.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name to be assigned to the workflow. It should be unique within your account.
- `Description` – UTF-8 string, not more than 120000 bytes long.

  A description of the workflow.
- `DefaultRunProperties` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a UTF-8 string.

  A collection of properties to be used as part of each execution of the workflow.

  Run properties may be logged. Do not pass plaintext secrets as properties. Retrieve secrets from a AWS Glue Connection, AWS Secrets Manager or other secret management mechanism if you intend to use them within the workflow run.
- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  The tags to be used with this workflow.
- `MaxConcurrentRuns` – Number (integer).

  You can use this parameter to prevent unwanted multiple updates to data, to control costs, or in some cases, to prevent exceeding the maximum number of concurrent runs of any of the component jobs. If you leave this parameter blank, there is no limit to the number of concurrent workflow runs.

Response

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the workflow which was provided as part of the request.

Errors

- `AlreadyExistsException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `ResourceNumberLimitExceededException`
- `ConcurrentModificationException`

UpdateWorkflow action (Python: update_workflow)

Updates an existing workflow.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Name of the workflow to be updated.

- `Description` – UTF-8 string, not more than 120000 bytes long.

  The description of the workflow.

- `DefaultRunProperties` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a UTF-8 string.

  A collection of properties to be used as part of each execution of the workflow.

Run properties may be logged. Do not pass plaintext secrets as properties. Retrieve secrets from a AWS Glue Connection, AWS Secrets Manager or other secret management mechanism if you intend to use them within the workflow run.

- `MaxConcurrentRuns` – Number (integer).

  You can use this parameter to prevent unwanted multiple updates to data, to control costs, or in some cases, to prevent exceeding the maximum number of concurrent runs of any of the component jobs. If you leave this parameter blank, there is no limit to the number of concurrent workflow runs.

Response

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the workflow which was specified in input.

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`
- `ConcurrentModificationException`

DeleteWorkflow action (Python: delete_workflow)

Deletes a workflow.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Name of the workflow to be deleted.

Response

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Name of the workflow specified in input.

Errors

- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `ConcurrentModificationException`

GetWorkflow action (Python: get_workflow)

Retrieves resource metadata for a workflow.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the workflow to retrieve.
- `IncludeGraph` – Boolean.

  Specifies whether to include a graph when returning the workflow resource metadata.

Response

- `Workflow` – A Workflow object.

  The resource metadata for the workflow.

Errors

- `InvalidInputException`
- `EntityNotFoundException`

- `InternalServiceException`
- `OperationTimeoutException`

ListWorkflows action (Python: list_workflows)

Lists names of workflows created in the account.

Request

- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation request.
- `MaxResults` – Number (integer), not less than 1 or more than 25.

  The maximum size of a list to return.

Response

- `Workflows` – An array of UTF-8 strings, not less than 1 or more than 25 strings.

  List of names of workflows in the account.
- `NextToken` – UTF-8 string.

  A continuation token, if not all workflow names have been returned.

Errors

- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`

BatchGetWorkflows action (Python: batch_get_workflows)

Returns a list of resource metadata for a given list of workflow names. After calling the `ListWorkflows` operation, you can call this operation to access the data to which you have been granted permissions. This operation supports all IAM permissions, including permission conditions that uses tags.

Request

- `Names` – *Required:* An array of UTF-8 strings, not less than 1 or more than 25 strings.

  A list of workflow names, which may be the names returned from the `ListWorkflows` operation.
- `IncludeGraph` – Boolean.

  Specifies whether to include a graph when returning the workflow resource metadata.

Response

- `Workflows` – An array of Workflow objects, not less than 1 or more than 25 structures.

  A list of workflow resource metadata.
- `MissingWorkflows` – An array of UTF-8 strings, not less than 1 or more than 25 strings.

  A list of names of workflows not found.

Errors

- `InternalServiceException`
- `OperationTimeoutException`
- `InvalidInputException`

GetWorkflowRun action (Python: get_workflow_run)

Retrieves the metadata for a given workflow run. Job run history is accessible for 90 days for your workflow and job run.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Name of the workflow being run.
- `RunId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The ID of the workflow run.

- `IncludeGraph` – Boolean.

  Specifies whether to include the workflow graph in response or not.

Response

- Run – A WorkflowRun object.

  The requested workflow run metadata.

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`

GetWorkflowRuns action (Python: get_workflow_runs)

Retrieves metadata for all runs of a given workflow.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Name of the workflow whose metadata of runs should be returned.
- `IncludeGraph` – Boolean.

  Specifies whether to include the workflow graph in response or not.
- `NextToken` – UTF-8 string.

  The maximum size of the response.
- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum number of workflow runs to be included in the response.

Response

- `Runs` – An array of [WorkflowRun](#) objects, not less than 1 or more than 1000 structures.

  A list of workflow run metadata objects.
- `NextToken` – UTF-8 string.

  A continuation token, if not all requested workflow runs have been returned.

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`

GetWorkflowRunProperties action (Python: get_workflow_run_properties)

Retrieves the workflow run properties which were set during the run.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Name of the workflow which was run.
- `RunId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the workflow run whose run properties should be returned.

Response

- `RunProperties` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

Each value is a UTF-8 string.

The workflow run properties which were set during the specified run.

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`

PutWorkflowRunProperties action (Python: put_workflow_run_properties)

Puts the specified workflow run properties for the given workflow run. If a property already exists for the specified run, then it overrides the value otherwise adds the property to existing properties.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Name of the workflow which was run.
- `RunId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the workflow run for which the run properties should be updated.
- `RunProperties` – *Required:* A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a UTF-8 string.

  The properties to put for the specified run.

Run properties may be logged. Do not pass plaintext secrets as properties. Retrieve secrets from a AWS Glue Connection, AWS Secrets Manager or other secret management mechanism if you intend to use them within the workflow run.

Response

- *No Response parameters.*

Errors

- `AlreadyExistsException`
- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `ResourceNumberLimitExceededException`
- `ConcurrentModificationException`

CreateBlueprint action (Python: create_blueprint)

Registers a blueprint with AWS Glue.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #31.

  The name of the blueprint.
- `Description` – UTF-8 string, not less than 1 or more than 512 bytes long.

  A description of the blueprint.
- `BlueprintLocation` – *Required:* UTF-8 string, not less than 1 or more than 8192 bytes long, matching the Custom string pattern #32.

  Specifies a path in Amazon S3 where the blueprint is published.

- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  The tags to be applied to this blueprint.

Response

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Returns the name of the blueprint that was registered.

Errors

- `AlreadyExistsException`
- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`
- `ResourceNumberLimitExceededException`

UpdateBlueprint action (Python: update_blueprint)

Updates a registered blueprint.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #31.

  The name of the blueprint.
- `Description` – UTF-8 string, not less than 1 or more than 512 bytes long.

  A description of the blueprint.
- `BlueprintLocation` – *Required:* UTF-8 string, not less than 1 or more than 8192 bytes long, matching the Custom string pattern #32.

Specifies a path in Amazon S3 where the blueprint is published.

Response

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Returns the name of the blueprint that was updated.

Errors

- `EntityNotFoundException`
- `ConcurrentModificationException`
- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`
- `IllegalBlueprintStateException`

DeleteBlueprint action (Python: delete_blueprint)

Deletes an existing blueprint.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the blueprint to delete.

Response

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Returns the name of the blueprint that was deleted.

Errors

- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`

ListBlueprints action (Python: list_blueprints)

Lists all the blueprint names in an account.

Request

- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation request.
- `MaxResults` – Number (integer), not less than 1 or more than 25.

  The maximum size of a list to return.
- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  Filters the list by an AWS resource tag.

Response

- `Blueprints` – An array of UTF-8 strings.

  List of names of blueprints in the account.
- `NextToken` – UTF-8 string.

  A continuation token, if not all blueprint names have been returned.

Errors

- `InvalidInputException`

- `InternalServiceException`
- `OperationTimeoutException`

BatchGetBlueprints action (Python: batch_get_blueprints)

Retrieves information about a list of blueprints.

Request

- `Names` – *Required:* An array of UTF-8 strings, not less than 1 or more than 25 strings.

  A list of blueprint names.
- `IncludeBlueprint` – Boolean.

  Specifies whether or not to include the blueprint in the response.
- `IncludeParameterSpec` – Boolean.

  Specifies whether or not to include the parameters, as a JSON string, for the blueprint in the response.

Response

- `Blueprints` – An array of Blueprint objects.

  Returns a list of blueprint as a `Blueprints` object.
- `MissingBlueprints` – An array of UTF-8 strings.

  Returns a list of `BlueprintNames` that were not found.

Errors

- `InternalServiceException`
- `OperationTimeoutException`
- `InvalidInputException`

StartBlueprintRun action (Python: start_blueprint_run)

Starts a new run of the specified blueprint.

Request

- `BlueprintName` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #31.

  The name of the blueprint.
- `Parameters` – UTF-8 string, not less than 1 or more than 131072 bytes long.

  Specifies the parameters as a `BlueprintParameters` object.
- `RoleArn` – *Required:* UTF-8 string, not less than 1 or more than 1024 bytes long, matching the Custom string pattern #30.

  Specifies the IAM role used to create the workflow.

Response

- `RunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The run ID for this blueprint run.

Errors

- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`
- `ResourceNumberLimitExceededException`
- `EntityNotFoundException`
- `IllegalBlueprintStateException`

GetBlueprintRun action (Python: get_blueprint_run)

Retrieves the details of a blueprint run.

Request

- `BlueprintName` – *Required:* UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #31.

  The name of the blueprint.
- `RunId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The run ID for the blueprint run you want to retrieve.

Response

- `BlueprintRun` – A BlueprintRun object.

  Returns a `BlueprintRun` object.

Errors

- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`

GetBlueprintRuns action (Python: get_blueprint_runs)

Retrieves the details of blueprint runs for a specified blueprint.

Request

- `BlueprintName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the blueprint.
- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation request.
- `MaxResults` – Number (integer), not less than 1 or more than 1000.

The maximum size of a list to return.

Response

- `BlueprintRuns` – An array of [BlueprintRun](#) objects.

  Returns a list of `BlueprintRun` objects.
- `NextToken` – UTF-8 string.

  A continuation token, if not all blueprint runs have been returned.

Errors

- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`
- `InvalidInputException`

StartWorkflowRun action (Python: start_workflow_run)

Starts a new run of the specified workflow.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the workflow to start.
- `RunProperties` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a UTF-8 string.

  The workflow run properties for the new workflow run.

Run properties may be logged. Do not pass plaintext secrets as properties. Retrieve secrets from a AWS Glue Connection, AWS Secrets Manager or other secret management mechanism if you intend to use them within the workflow run.

Response

- RunId – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  An Id for the new run.

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`
- `ResourceNumberLimitExceededException`
- `ConcurrentRunsExceededException`

StopWorkflowRun action (Python: stop_workflow_run)

Stops the execution of the specified workflow run.

Request

- Name – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the workflow to stop.

- RunId – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the workflow run to stop.

Response

- *No Response parameters.*

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`
- `IllegalWorkflowStateException`

ResumeWorkflowRun action (Python: resume_workflow_run)

Restarts selected nodes of a previous partially completed workflow run and resumes the workflow run. The selected nodes and all nodes that are downstream from the selected nodes are run.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the workflow to resume.
- `RunId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the workflow run to resume.
- `NodeIds` – *Required:* An array of UTF-8 strings.

  A list of the node IDs for the nodes you want to restart. The nodes that are to be restarted must have a run attempt in the original run.

Response

- `RunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The new ID assigned to the resumed workflow run. Each resume of a workflow run will have a new run ID.

- `NodeIds` – An array of UTF-8 strings.

  A list of the node IDs for the nodes that were actually restarted.

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `InternalServiceException`
- `OperationTimeoutException`
- `ConcurrentRunsExceededException`
- `IllegalWorkflowStateException`

Usage profiles

The Usage profiles API describes the data types and API related to creating, updating, or viewing usage profiles in AWS Glue.

Data types

- ProfileConfiguration structure
- ConfigurationObject structure
- UsageProfileDefinition structure

ProfileConfiguration structure

Specifies the job and session values that an admin configures in an AWS Glue usage profile.

Fields

- `SessionConfiguration` – A map array of key-value pairs.

Each key is a UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

Each value is a A ConfigurationObject object.

A key-value map of configuration parameters for AWS Glue sessions.

- `JobConfiguration` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a A ConfigurationObject object.

  A key-value map of configuration parameters for AWS Glue jobs.

ConfigurationObject structure

Specifies the values that an admin sets for each job or session parameter configured in a AWS Glue usage profile.

Fields

- `DefaultValue` – UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #35.

  A default value for the parameter.
- `AllowedValues` – An array of UTF-8 strings.

  A list of allowed values for the parameter.
- `MinValue` – UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #35.

  A minimum allowed value for the parameter.
- `MaxValue` – UTF-8 string, not less than 1 or more than 128 bytes long, matching the Custom string pattern #35.

  A maximum allowed value for the parameter.

UsageProfileDefinition structure

Describes an AWS Glue usage profile.

Fields

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the usage profile.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the usage profile.
- `CreatedOn` – Timestamp.

  The date and time when the usage profile was created.
- `LastModifiedOn` – Timestamp.

  The date and time when the usage profile was last modified.

Operations

- CreateUsageProfile action (Python: create_usage_profile)
- GetUsageProfile action (Python: get_usage_profile)
- UpdateUsageProfile action (Python: update_usage_profile)
- DeleteUsageProfile action (Python: delete_usage_profile)
- ListUsageProfiles action (Python: list_usage_profiles)

CreateUsageProfile action (Python: create_usage_profile)

Creates an AWS Glue usage profile.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The name of the usage profile.

- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the usage profile.
- `Configuration` – *Required:* A ProfileConfiguration object.

  A `ProfileConfiguration` object specifying the job and session values for the profile.
- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  A list of tags applied to the usage profile.

Response

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the usage profile that was created.

Errors

- `InvalidInputException`
- `InternalServiceException`
- `AlreadyExistsException`
- `OperationTimeoutException`
- `ResourceNumberLimitExceededException`
- `OperationNotSupportedException`

GetUsageProfile action (Python: get_usage_profile)

Retrieves information about the specified AWS Glue usage profile.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the usage profile to retrieve.

Response

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the usage profile.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the usage profile.
- `Configuration` – A ProfileConfiguration object.

  A `ProfileConfiguration` object specifying the job and session values for the profile.
- `CreatedOn` – Timestamp.

  The date and time when the usage profile was created.
- `LastModifiedOn` – Timestamp.

  The date and time when the usage profile was last modified.

Errors

- `InvalidInputException`
- `InternalServiceException`
- `EntityNotFoundException`
- `OperationTimeoutException`
- `OperationNotSupportedException`

UpdateUsageProfile action (Python: update_usage_profile)

Update an AWS Glue usage profile.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the usage profile.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the usage profile.
- `Configuration` – *Required:* A ProfileConfiguration object.

  A `ProfileConfiguration` object specifying the job and session values for the profile.

Response

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the usage profile that was updated.

Errors

- `InvalidInputException`
- `InternalServiceException`
- `EntityNotFoundException`
- `OperationTimeoutException`
- `OperationNotSupportedException`
- `ConcurrentModificationException`

DeleteUsageProfile action (Python: delete_usage_profile)

Deletes the AWS Glue specified usage profile.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the [Single-line string pattern](#).

  The name of the usage profile to delete.

Response

- *No Response parameters.*

Errors

- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `OperationNotSupportedException`

ListUsageProfiles action (Python: list_usage_profiles)

List all the AWS Glue usage profiles.

Request

- `NextToken` – UTF-8 string, not more than 400000 bytes long.

  A continuation token, included if this is a continuation call.
- `MaxResults` – Number (integer), not less than 1 or more than 200.

  The maximum number of usage profiles to return in a single response.

Response

- `Profiles` – An array of [UsageProfileDefinition](#) objects.

  A list of usage profile (`UsageProfileDefinition`) objects.
- `NextToken` – UTF-8 string, not more than 400000 bytes long.

A continuation token, present if the current list segment is not the last.

Errors

- `InternalServiceException`
- `OperationTimeoutException`
- `InvalidInputException`
- `OperationNotSupportedException`

Machine learning API

The Machine learning API describes the machine learning data types, and includes the API for creating, deleting, or updating a transform, or starting a machine learning task run.

Data types

- TransformParameters structure
- EvaluationMetrics structure
- MLTransform structure
- FindMatchesParameters structure
- FindMatchesMetrics structure
- ConfusionMatrix structure
- GlueTable structure
- TaskRun structure
- TransformFilterCriteria structure
- TransformSortCriteria structure
- TaskRunFilterCriteria structure
- TaskRunSortCriteria structure
- TaskRunProperties structure
- FindMatchesTaskRunProperties structure
- ImportLabelsTaskRunProperties structure
- ExportLabelsTaskRunProperties structure

- LabelingSetGenerationTaskRunProperties structure
- SchemaColumn structure
- TransformEncryption structure
- MLUserDataEncryption structure
- ColumnImportance structure

TransformParameters structure

The algorithm-specific parameters that are associated with the machine learning transform.

Fields

- `TransformType` – *Required:* UTF-8 string (valid values: `FIND_MATCHES`).

  The type of machine learning transform.

  For information about the types of machine learning transforms, see Creating Machine Learning Transforms.
- `FindMatchesParameters` – A FindMatchesParameters object.

  The parameters for the find matches algorithm.

EvaluationMetrics structure

Evaluation metrics provide an estimate of the quality of your machine learning transform.

Fields

- `TransformType` – *Required:* UTF-8 string (valid values: `FIND_MATCHES`).

  The type of machine learning transform.
- `FindMatchesMetrics` – A FindMatchesMetrics object.

  The evaluation metrics for the find matches algorithm.

MLTransform structure

A structure for a machine learning transform.

Fields

- `TransformId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique transform ID that is generated for the machine learning transform. The ID is guaranteed to be unique and does not change.
- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A user-defined name for the machine learning transform. Names are not guaranteed unique and can be changed at any time.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A user-defined, long-form description text for the machine learning transform. Descriptions are not guaranteed to be unique and can be changed at any time.
- `Status` – UTF-8 string (valid values: `NOT_READY` | `READY` | `DELETING`).

  The current status of the machine learning transform.
- `CreatedOn` – Timestamp.

  A timestamp. The time and date that this machine learning transform was created.
- `LastModifiedOn` – Timestamp.

  A timestamp. The last point in time when this machine learning transform was modified.
- `InputRecordTables` – An array of GlueTable objects, not more than 10 structures.

  A list of AWS Glue table definitions used by the transform.
- `Parameters` – A TransformParameters object.

  A `TransformParameters` object. You can use parameters to tune (customize) the behavior of the machine learning transform by specifying what data it learns from and your preference on various tradeoffs (such as precious vs. recall, or accuracy vs. cost).
- `EvaluationMetrics` – An EvaluationMetrics object.

  An `EvaluationMetrics` object. Evaluation metrics provide an estimate of the quality of your machine learning transform.

- `LabelCount` – Number (integer).

  A count identifier for the labeling files generated by AWS Glue for this transform. As you create a better transform, you can iteratively download, label, and upload the labeling file.
- `Schema` – An array of SchemaColumn objects, not more than 100 structures.

  A map of key-value pairs representing the columns and data types that this transform can run against. Has an upper bound of 100 columns.
- `Role` – UTF-8 string.

  The name or Amazon Resource Name (ARN) of the IAM role with the required permissions. The required permissions include both AWS Glue service role permissions to AWS Glue resources, and Amazon S3 permissions required by the transform.
  - This role needs AWS Glue service role permissions to allow access to resources in AWS Glue. See Attach a Policy to IAM Users That Access AWS Glue.
  - This role needs permission to your Amazon Simple Storage Service (Amazon S3) sources, targets, temporary directory, scripts, and any libraries used by the task run for this transform.
- `GlueVersion` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #47.

  This value determines which version of AWS Glue this machine learning transform is compatible with. Glue 1.0 is recommended for most customers. If the value is not set, the Glue compatibility defaults to Glue 0.9. For more information, see AWS Glue Versions in the developer guide.
- `MaxCapacity` – Number (double).

  The number of AWS Glue data processing units (DPUs) that are allocated to task runs for this transform. You can allocate from 2 to 100 DPUs; the default is 10. A DPU is a relative measure of processing power that consists of 4 vCPUs of compute capacity and 16 GB of memory. For more information, see the AWS Glue pricing page.

  `MaxCapacity` is a mutually exclusive option with `NumberOfWorkers` and `WorkerType`.
  - If either `NumberOfWorkers` or `WorkerType` is set, then `MaxCapacity` cannot be set.
  - If `MaxCapacity` is set then neither `NumberOfWorkers` or `WorkerType` can be set.
  - If `WorkerType` is set, then `NumberOfWorkers` is required (and vice versa).
  - `MaxCapacity` and `NumberOfWorkers` must both be at least 1.

When the `WorkerType` field is set to a value other than `Standard`, the `MaxCapacity` field is set automatically and becomes read-only.

- `WorkerType` – UTF-8 string (valid values: `Standard=""` | `G.1X=""` | `G.2X=""` | `G.025X=""` | `G.4X=""` | `G.8X=""` | `Z.2X=""`).

  The type of predefined worker that is allocated when a task of this transform runs. Accepts a value of Standard, G.1X, or G.2X.

  - For the `Standard` worker type, each worker provides 4 vCPU, 16 GB of memory and a 50GB disk, and 2 executors per worker.
  - For the `G.1X` worker type, each worker provides 4 vCPU, 16 GB of memory and a 64GB disk, and 1 executor per worker.
  - For the `G.2X` worker type, each worker provides 8 vCPU, 32 GB of memory and a 128GB disk, and 1 executor per worker.

  `MaxCapacity` is a mutually exclusive option with `NumberOfWorkers` and `WorkerType`.

  - If either `NumberOfWorkers` or `WorkerType` is set, then `MaxCapacity` cannot be set.
  - If `MaxCapacity` is set then neither `NumberOfWorkers` or `WorkerType` can be set.
  - If `WorkerType` is set, then `NumberOfWorkers` is required (and vice versa).
  - `MaxCapacity` and `NumberOfWorkers` must both be at least 1.
- `NumberOfWorkers` – Number (integer).

  The number of workers of a defined `workerType` that are allocated when a task of the transform runs.

  If `WorkerType` is set, then `NumberOfWorkers` is required (and vice versa).
- `Timeout` – Number (integer), at least 1.

  The timeout in minutes of the machine learning transform.
- `MaxRetries` – Number (integer).

  The maximum number of times to retry after an `MLTaskRun` of the machine learning transform fails.
- `TransformEncryption` – A TransformEncryption object.

  The encryption-at-rest settings of the transform that apply to accessing user data. Machine learning transforms can access user data encrypted in Amazon S3 using KMS.

FindMatchesParameters structure

The parameters to configure the find matches transform.

Fields

- `PrimaryKeyColumnName` – UTF-8 string, not less than 1 or more than 1024 bytes long, matching the Single-line string pattern.

  The name of a column that uniquely identifies rows in the source table. Used to help identify matching records.
- `PrecisionRecallTradeoff` – Number (double), not more than 1.0.

  The value selected when tuning your transform for a balance between precision and recall. A value of 0.5 means no preference; a value of 1.0 means a bias purely for precision, and a value of 0.0 means a bias for recall. Because this is a tradeoff, choosing values close to 1.0 means very low recall, and choosing values close to 0.0 results in very low precision.

  The precision metric indicates how often your model is correct when it predicts a match.

  The recall metric indicates that for an actual match, how often your model predicts the match.
- `AccuracyCostTradeoff` – Number (double), not more than 1.0.

  The value that is selected when tuning your transform for a balance between accuracy and cost. A value of 0.5 means that the system balances accuracy and cost concerns. A value of 1.0 means a bias purely for accuracy, which typically results in a higher cost, sometimes substantially higher. A value of 0.0 means a bias purely for cost, which results in a less accurate `FindMatches` transform, sometimes with unacceptable accuracy.

  Accuracy measures how well the transform finds true positives and true negatives. Increasing accuracy requires more machine resources and cost. But it also results in increased recall.

  Cost measures how many compute resources, and thus money, are consumed to run the transform.
- `EnforceProvidedLabels` – Boolean.

  The value to switch on or off to force the output to match the provided labels from users. If the value is `True`, the `find matches` transform forces the output to match the provided labels. The results override the normal conflation results. If the value is `False`, the `find matches` transform does not ensure all the labels provided are respected, and the results rely on the trained model.

Note that setting this value to true may increase the conflation execution time.

FindMatchesMetrics structure

The evaluation metrics for the find matches algorithm. The quality of your machine learning transform is measured by getting your transform to predict some matches and comparing the results to known matches from the same dataset. The quality metrics are based on a subset of your data, so they are not precise.

Fields

- `AreaUnderPRCurve` – Number (double), not more than 1.0.

  The area under the precision/recall curve (AUPRC) is a single number measuring the overall quality of the transform, that is independent of the choice made for precision vs. recall. Higher values indicate that you have a more attractive precision vs. recall tradeoff.

  For more information, see Precision and recall in Wikipedia.
- `Precision` – Number (double), not more than 1.0.

  The precision metric indicates when often your transform is correct when it predicts a match. Specifically, it measures how well the transform finds true positives from the total true positives possible.

  For more information, see Precision and recall in Wikipedia.
- `Recall` – Number (double), not more than 1.0.

  The recall metric indicates that for an actual match, how often your transform predicts the match. Specifically, it measures how well the transform finds true positives from the total records in the source data.

  For more information, see Precision and recall in Wikipedia.
- `F1` – Number (double), not more than 1.0.

  The maximum F1 metric indicates the transform's accuracy between 0 and 1, where 1 is the best accuracy.

For more information, see F1 score in Wikipedia.

- `ConfusionMatrix` – A ConfusionMatrix object.

  The confusion matrix shows you what your transform is predicting accurately and what types of errors it is making.

  For more information, see Confusion matrix in Wikipedia.
- `ColumnImportances` – An array of ColumnImportance objects, not more than 100 structures.

  A list of `ColumnImportance` structures containing column importance metrics, sorted in order of descending importance.

ConfusionMatrix structure

The confusion matrix shows you what your transform is predicting accurately and what types of errors it is making.

For more information, see Confusion matrix in Wikipedia.

Fields

- `NumTruePositives` – Number (long).

  The number of matches in the data that the transform correctly found, in the confusion matrix for your transform.
- `NumFalsePositives` – Number (long).

  The number of nonmatches in the data that the transform incorrectly classified as a match, in the confusion matrix for your transform.
- `NumTrueNegatives` – Number (long).

  The number of nonmatches in the data that the transform correctly rejected, in the confusion matrix for your transform.
- `NumFalseNegatives` – Number (long).

  The number of matches in the data that the transform didn't find, in the confusion matrix for your transform.

GlueTable structure

The database and table in the AWS Glue Data Catalog that is used for input or output data.

Fields

- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A database name in the AWS Glue Data Catalog.
- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A table name in the AWS Glue Data Catalog.
- `CatalogId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A unique identifier for the AWS Glue Data Catalog.
- `ConnectionName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the connection to the AWS Glue Data Catalog.
- `AdditionalOptions` – A map array of key-value pairs, not less than 1 or more than 10 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  Additional options for the table. Currently there are two keys supported:
  - `pushDownPredicate`: to filter on partitions without having to list and read all the files in your dataset.
  - `catalogPartitionPredicate`: to use server-side partition pruning using partition indexes in the AWS Glue Data Catalog.

TaskRun structure

The sampling parameters that are associated with the machine learning transform.

Fields

- `TransformId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique identifier for the transform.
- `TaskRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique identifier for this task run.
- `Status` – UTF-8 string (valid values: `RUNNING` | `FINISHED` | `FAILED` | `PENDING_EXECUTION` | `TIMED_OUT` | `CANCELING` | `CANCELED` | `RECEIVED_BY_TASKRUNNER`).

  The current status of the requested task run.
- `LogGroupName` – UTF-8 string.

  The names of the log group for secure logging, associated with this task run.
- `Properties` – A TaskRunProperties object.

  Specifies configuration properties associated with this task run.
- `ErrorString` – UTF-8 string.

  The list of error strings associated with this task run.
- `StartedOn` – Timestamp.

  The date and time that this task run started.
- `LastModifiedOn` – Timestamp.

  The last point in time that the requested task run was updated.
- `CompletedOn` – Timestamp.

  The last point in time that the requested task run was completed.
- `ExecutionTime` – Number (integer).

  The amount of time (in seconds) that the task run consumed resources.

TransformFilterCriteria structure

The criteria used to filter the machine learning transforms.

Fields

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A unique transform name that is used to filter the machine learning transforms.
- `TransformType` – UTF-8 string (valid values: `FIND_MATCHES`).

  The type of machine learning transform that is used to filter the machine learning transforms.
- `Status` – UTF-8 string (valid values: `NOT_READY` | `READY` | `DELETING`).

  Filters the list of machine learning transforms by the last known status of the transforms (to indicate whether a transform can be used or not). One of "NOT_READY", "READY", or "DELETING".
- `GlueVersion` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #47.

  This value determines which version of AWS Glue this machine learning transform is compatible with. Glue 1.0 is recommended for most customers. If the value is not set, the Glue compatibility defaults to Glue 0.9. For more information, see AWS Glue Versions in the developer guide.
- `CreatedBefore` – Timestamp.

  The time and date before which the transforms were created.
- `CreatedAfter` – Timestamp.

  The time and date after which the transforms were created.
- `LastModifiedBefore` – Timestamp.

  Filter on transforms last modified before this date.
- `LastModifiedAfter` – Timestamp.

  Filter on transforms last modified after this date.
- `Schema` – An array of SchemaColumn objects, not more than 100 structures.

  Filters on datasets with a specific schema. The `Map<Column, Type>` object is an array of key-value pairs representing the schema this transform accepts, where `Column` is the name of a column, and `Type` is the type of the data such as an integer or string. Has an upper bound of 100 columns.

TransformSortCriteria structure

The sorting criteria that are associated with the machine learning transform.

Fields

- `Column` – *Required:* UTF-8 string (valid values: `NAME` | `TRANSFORM_TYPE` | `STATUS` | `CREATED` | `LAST_MODIFIED`).

  The column to be used in the sorting criteria that are associated with the machine learning transform.
- `SortDirection` – *Required:* UTF-8 string (valid values: `DESCENDING` | `ASCENDING`).

  The sort direction to be used in the sorting criteria that are associated with the machine learning transform.

TaskRunFilterCriteria structure

The criteria that are used to filter the task runs for the machine learning transform.

Fields

- `TaskRunType` – UTF-8 string (valid values: `EVALUATION` | `LABELING_SET_GENERATION` | `IMPORT_LABELS` | `EXPORT_LABELS` | `FIND_MATCHES`).

  The type of task run.
- `Status` – UTF-8 string (valid values: `RUNNING` | `FINISHED` | `FAILED` | `PENDING_EXECUTION` | `TIMED_OUT` | `CANCELING` | `CANCELED` | `RECEIVED_BY_TASKRUNNER`).

  The current status of the task run.
- `StartedBefore` – Timestamp.

  Filter on task runs started before this date.
- `StartedAfter` – Timestamp.

  Filter on task runs started after this date.

TaskRunSortCriteria structure

The sorting criteria that are used to sort the list of task runs for the machine learning transform.

Fields

- `Column` – *Required:* UTF-8 string (valid values: `TASK_RUN_TYPE` | `STATUS` | `STARTED`).

  The column to be used to sort the list of task runs for the machine learning transform.
- `SortDirection` – *Required:* UTF-8 string (valid values: `DESCENDING` | `ASCENDING`).

  The sort direction to be used to sort the list of task runs for the machine learning transform.

TaskRunProperties structure

The configuration properties for the task run.

Fields

- `TaskType` – UTF-8 string (valid values: `EVALUATION` | `LABELING_SET_GENERATION` | `IMPORT_LABELS` | `EXPORT_LABELS` | `FIND_MATCHES`).

  The type of task run.
- `ImportLabelsTaskRunProperties` – An ImportLabelsTaskRunProperties object.

  The configuration properties for an importing labels task run.
- `ExportLabelsTaskRunProperties` – An ExportLabelsTaskRunProperties object.

  The configuration properties for an exporting labels task run.
- `LabelingSetGenerationTaskRunProperties` – A LabelingSetGenerationTaskRunProperties object.

  The configuration properties for a labeling set generation task run.
- `FindMatchesTaskRunProperties` – A FindMatchesTaskRunProperties object.

  The configuration properties for a find matches task run.

FindMatchesTaskRunProperties structure

Specifies configuration properties for a Find Matches task run.

Fields

- `JobId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The job ID for the Find Matches task run.
- `JobName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name assigned to the job for the Find Matches task run.
- `JobRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The job run ID for the Find Matches task run.

ImportLabelsTaskRunProperties structure

Specifies configuration properties for an importing labels task run.

Fields

- `InputS3Path` – UTF-8 string.

  The Amazon Simple Storage Service (Amazon S3) path from where you will import the labels.
- `Replace` – Boolean.

  Indicates whether to overwrite your existing labels.

ExportLabelsTaskRunProperties structure

Specifies configuration properties for an exporting labels task run.

Fields

- `OutputS3Path` – UTF-8 string.

The Amazon Simple Storage Service (Amazon S3) path where you will export the labels.

LabelingSetGenerationTaskRunProperties structure

Specifies configuration properties for a labeling set generation task run.

Fields

- `OutputS3Path` – UTF-8 string.

  The Amazon Simple Storage Service (Amazon S3) path where you will generate the labeling set.

SchemaColumn structure

A key-value pair representing a column and data type that this transform can run against. The `Schema` parameter of the `MLTransform` may contain up to 100 of these structures.

Fields

- `Name` – UTF-8 string, not less than 1 or more than 1024 bytes long, matching the Single-line string pattern.

  The name of the column.
- `DataType` – UTF-8 string, not more than 131072 bytes long, matching the Single-line string pattern.

  The type of data in the column.

TransformEncryption structure

The encryption-at-rest settings of the transform that apply to accessing user data. Machine learning transforms can access user data encrypted in Amazon S3 using KMS.

Additionally, imported labels and trained transforms can now be encrypted using a customer provided KMS key.

Fields

- `MlUserDataEncryption` – A MLUserDataEncryption object.

An `MLUserDataEncryption` object containing the encryption mode and customer-provided KMS key ID.

- `TaskRunSecurityConfigurationName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the security configuration.

MLUserDataEncryption structure

The encryption-at-rest settings of the transform that apply to accessing user data.

Fields

- `MlUserDataEncryptionMode` – *Required:* UTF-8 string (valid values: `DISABLED` | `SSE-KMS="SSEKMS"`).

  The encryption mode applied to user data. Valid values are:

  - DISABLED: encryption is disabled
  - SSEKMS: use of server-side encryption with AWS Key Management Service (SSE-KMS) for user data stored in Amazon S3.
- `KmsKeyId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID for the customer-provided KMS key.

ColumnImportance structure

A structure containing the column name and column importance score for a column.

Column importance helps you understand how columns contribute to your model, by identifying which columns in your records are more important than others.

Fields

- `ColumnName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of a column.

- `Importance` – Number (double), not more than 1.0.

  The column importance score for the column, as a decimal.

Operations

- CreateMLTransform action (Python: create_ml_transform)
- UpdateMLTransform action (Python: update_ml_transform)
- DeleteMLTransform action (Python: delete_ml_transform)
- GetMLTransform action (Python: get_ml_transform)
- GetMLTransforms action (Python: get_ml_transforms)
- ListMLTransforms action (Python: list_ml_transforms)
- StartMLEvaluationTaskRun action (Python: start_ml_evaluation_task_run)
- StartMLLabelingSetGenerationTaskRun action (Python: start_ml_labeling_set_generation_task_run)
- GetMLTaskRun action (Python: get_ml_task_run)
- GetMLTaskRuns action (Python: get_ml_task_runs)
- CancelMLTaskRun action (Python: cancel_ml_task_run)
- StartExportLabelsTaskRun action (Python: start_export_labels_task_run)
- StartImportLabelsTaskRun action (Python: start_import_labels_task_run)

CreateMLTransform action (Python: create_ml_transform)

Creates an AWS Glue machine learning transform. This operation creates the transform and all the necessary parameters to train it.

Call this operation as the first step in the process of using a machine learning transform (such as the `FindMatches` transform) for deduplicating data. You can provide an optional `Description`, in addition to the parameters that you want to use for your algorithm.

You must also specify certain parameters for the tasks that AWS Glue runs on your behalf as part of learning from your data and creating a high-quality machine learning transform. These parameters include `Role`, and optionally, `AllocatedCapacity`, `Timeout`, and `MaxRetries`. For more information, see Jobs.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique name that you give the transform when you create it.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the machine learning transform that is being defined. The default is an empty string.
- `InputRecordTables` – *Required:* An array of GlueTable objects, not more than 10 structures.

  A list of AWS Glue table definitions used by the transform.
- `Parameters` – *Required:* A TransformParameters object.

  The algorithmic parameters that are specific to the transform type used. Conditionally dependent on the transform type.
- `Role` – *Required:* UTF-8 string.

  The name or Amazon Resource Name (ARN) of the IAM role with the required permissions. The required permissions include both AWS Glue service role permissions to AWS Glue resources, and Amazon S3 permissions required by the transform.
  - This role needs AWS Glue service role permissions to allow access to resources in AWS Glue. See Attach a Policy to IAM Users That Access AWS Glue.
  - This role needs permission to your Amazon Simple Storage Service (Amazon S3) sources, targets, temporary directory, scripts, and any libraries used by the task run for this transform.
- `GlueVersion` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #47.

  This value determines which version of AWS Glue this machine learning transform is compatible with. Glue 1.0 is recommended for most customers. If the value is not set, the Glue compatibility defaults to Glue 0.9. For more information, see AWS Glue Versions in the developer guide.
- `MaxCapacity` – Number (double).

  The number of AWS Glue data processing units (DPUs) that are allocated to task runs for this transform. You can allocate from 2 to 100 DPUs; the default is 10. A DPU is a relative measure of processing power that consists of 4 vCPUs of compute capacity and 16 GB of memory. For more information, see the [AWS Glue pricing page](#).

`MaxCapacity` is a mutually exclusive option with `NumberOfWorkers` and `WorkerType`.

- If either `NumberOfWorkers` or `WorkerType` is set, then `MaxCapacity` cannot be set.
- If `MaxCapacity` is set then neither `NumberOfWorkers` or `WorkerType` can be set.
- If `WorkerType` is set, then `NumberOfWorkers` is required (and vice versa).
- `MaxCapacity` and `NumberOfWorkers` must both be at least 1.

When the `WorkerType` field is set to a value other than `Standard`, the `MaxCapacity` field is set automatically and becomes read-only.

When the `WorkerType` field is set to a value other than `Standard`, the `MaxCapacity` field is set automatically and becomes read-only.

- `WorkerType` – UTF-8 string (valid values: `Standard=""` | `G.1X=""` | `G.2X=""` | `G.025X=""` | `G.4X=""` | `G.8X=""` | `Z.2X=""`).

  The type of predefined worker that is allocated when this task runs. Accepts a value of Standard, G.1X, or G.2X.

  - For the `Standard` worker type, each worker provides 4 vCPU, 16 GB of memory and a 50GB disk, and 2 executors per worker.
  - For the `G.1X` worker type, each worker provides 4 vCPU, 16 GB of memory and a 64GB disk, and 1 executor per worker.
  - For the `G.2X` worker type, each worker provides 8 vCPU, 32 GB of memory and a 128GB disk, and 1 executor per worker.

  `MaxCapacity` is a mutually exclusive option with `NumberOfWorkers` and `WorkerType`.

  - If either `NumberOfWorkers` or `WorkerType` is set, then `MaxCapacity` cannot be set.
  - If `MaxCapacity` is set then neither `NumberOfWorkers` or `WorkerType` can be set.
  - If `WorkerType` is set, then `NumberOfWorkers` is required (and vice versa).
  - `MaxCapacity` and `NumberOfWorkers` must both be at least 1.
- `NumberOfWorkers` – Number (integer).

  The number of workers of a defined `workerType` that are allocated when this task runs.

  If `WorkerType` is set, then `NumberOfWorkers` is required (and vice versa).

- `Timeout` – Number (integer), at least 1.

  The timeout of the task run for this transform in minutes. This is the maximum time that a task run for this transform can consume resources before it is terminated and enters `TIMEOUT` status. The default is 2,880 minutes (48 hours).
- `MaxRetries` – Number (integer).

  The maximum number of times to retry a task for this transform after a task run fails.
- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  The tags to use with this machine learning transform. You may use tags to limit access to the machine learning transform. For more information about tags in AWS Glue, see AWS Tags in AWS Glue in the developer guide.
- `TransformEncryption` – A TransformEncryption object.

  The encryption-at-rest settings of the transform that apply to accessing user data. Machine learning transforms can access user data encrypted in Amazon S3 using KMS.

Response

- `TransformId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A unique identifier that is generated for the transform.

Errors

- `AlreadyExistsException`
- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`
- `AccessDeniedException`
- `ResourceNumberLimitExceededException`

- `IdempotentParameterMismatchException`

UpdateMLTransform action (Python: update_ml_transform)

Updates an existing machine learning transform. Call this operation to tune the algorithm parameters to achieve better results.

After calling this operation, you can call the `StartMLEvaluationTaskRun` operation to assess how well your new parameters achieved your goals (such as improving the quality of your machine learning transform, or making it more cost-effective).

Request

- `TransformId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A unique identifier that was generated when the transform was created.
- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique name that you gave the transform when you created it.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the transform. The default is an empty string.
- `Parameters` – A TransformParameters object.

  The configuration parameters that are specific to the transform type (algorithm) used. Conditionally dependent on the transform type.
- `Role` – UTF-8 string.

  The name or Amazon Resource Name (ARN) of the IAM role with the required permissions.
- `GlueVersion` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #47.

  This value determines which version of AWS Glue this machine learning transform is compatible with. Glue 1.0 is recommended for most customers. If the value is not set, the Glue compatibility defaults to Glue 0.9. For more information, see AWS Glue Versions in the developer guide.

- `MaxCapacity` – Number (double).

  The number of AWS Glue data processing units (DPUs) that are allocated to task runs for this transform. You can allocate from 2 to 100 DPUs; the default is 10. A DPU is a relative measure of processing power that consists of 4 vCPUs of compute capacity and 16 GB of memory. For more information, see the AWS Glue pricing page.

  When the `WorkerType` field is set to a value other than `Standard`, the `MaxCapacity` field is set automatically and becomes read-only.
- `WorkerType` – UTF-8 string (valid values: `Standard=""` | `G.1X=""` | `G.2X=""` | `G.025X=""` | `G.4X=""` | `G.8X=""` | `Z.2X=""`).

  The type of predefined worker that is allocated when this task runs. Accepts a value of Standard, G.1X, or G.2X.
  - For the `Standard` worker type, each worker provides 4 vCPU, 16 GB of memory and a 50GB disk, and 2 executors per worker.
  - For the `G.1X` worker type, each worker provides 4 vCPU, 16 GB of memory and a 64GB disk, and 1 executor per worker.
  - For the `G.2X` worker type, each worker provides 8 vCPU, 32 GB of memory and a 128GB disk, and 1 executor per worker.
- `NumberOfWorkers` – Number (integer).

  The number of workers of a defined `workerType` that are allocated when this task runs.
- `Timeout` – Number (integer), at least 1.

  The timeout for a task run for this transform in minutes. This is the maximum time that a task run for this transform can consume resources before it is terminated and enters `TIMEOUT` status. The default is 2,880 minutes (48 hours).
- `MaxRetries` – Number (integer).

  The maximum number of times to retry a task for this transform after a task run fails.

Response

- `TransformId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The unique identifier for the transform that was updated.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`
- `AccessDeniedException`

DeleteMLTransform action (Python: delete_ml_transform)

Deletes an AWS Glue machine learning transform. Machine learning transforms are a special type of transform that use machine learning to learn the details of the transformation to be performed by learning from examples provided by humans. These transformations are then saved by AWS Glue. If you no longer need a transform, you can delete it by calling `DeleteMLTransforms`. However, any AWS Glue jobs that still reference the deleted transform will no longer succeed.

Request

- `TransformId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique identifier of the transform to delete.

Response

- `TransformId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique identifier of the transform that was deleted.

Errors

- `EntityNotFoundException`
- `InvalidInputException`

- `OperationTimeoutException`
- `InternalServiceException`

GetMLTransform action (Python: get_ml_transform)

Gets an AWS Glue machine learning transform artifact and all its corresponding metadata. Machine learning transforms are a special type of transform that use machine learning to learn the details of the transformation to be performed by learning from examples provided by humans. These transformations are then saved by AWS Glue. You can retrieve their metadata by calling `GetMLTransform`.

Request

- `TransformId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique identifier of the transform, generated at the time that the transform was created.

Response

- `TransformId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique identifier of the transform, generated at the time that the transform was created.
- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique name given to the transform when it was created.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the transform.
- `Status` – UTF-8 string (valid values: `NOT_READY` | `READY` | `DELETING`).

  The last known status of the transform (to indicate whether it can be used or not). One of "NOT_READY", "READY", or "DELETING".
- `CreatedOn` – Timestamp.

The date and time when the transform was created.

- `LastModifiedOn` – Timestamp.

  The date and time when the transform was last modified.

- `InputRecordTables` – An array of GlueTable objects, not more than 10 structures.

  A list of AWS Glue table definitions used by the transform.

- `Parameters` – A TransformParameters object.

  The configuration parameters that are specific to the algorithm used.

- `EvaluationMetrics` – An EvaluationMetrics object.

  The latest evaluation metrics.

- `LabelCount` – Number (integer).

  The number of labels available for this transform.

- `Schema` – An array of SchemaColumn objects, not more than 100 structures.

  The `Map<Column, Type>` object that represents the schema that this transform accepts. Has an upper bound of 100 columns.

- `Role` – UTF-8 string.

  The name or Amazon Resource Name (ARN) of the IAM role with the required permissions.

- `GlueVersion` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #47.

  This value determines which version of AWS Glue this machine learning transform is compatible with. Glue 1.0 is recommended for most customers. If the value is not set, the Glue compatibility defaults to Glue 0.9. For more information, see AWS Glue Versions in the developer guide.

- `MaxCapacity` – Number (double).

  The number of AWS Glue data processing units (DPUs) that are allocated to task runs for this transform. You can allocate from 2 to 100 DPUs; the default is 10. A DPU is a relative measure of processing power that consists of 4 vCPUs of compute capacity and 16 GB of memory. For more information, see the AWS Glue pricing page.

When the `WorkerType` field is set to a value other than `Standard`, the `MaxCapacity` field is set automatically and becomes read-only.

- `WorkerType` – UTF-8 string (valid values: `Standard=""` | `G.1X=""` | `G.2X=""` | `G.025X=""` | `G.4X=""` | `G.8X=""` | `Z.2X=""`).

  The type of predefined worker that is allocated when this task runs. Accepts a value of Standard, G.1X, or G.2X.

  - For the `Standard` worker type, each worker provides 4 vCPU, 16 GB of memory and a 50GB disk, and 2 executors per worker.
  - For the `G.1X` worker type, each worker provides 4 vCPU, 16 GB of memory and a 64GB disk, and 1 executor per worker.
  - For the `G.2X` worker type, each worker provides 8 vCPU, 32 GB of memory and a 128GB disk, and 1 executor per worker.
- `NumberOfWorkers` – Number (integer).

  The number of workers of a defined `workerType` that are allocated when this task runs.
- `Timeout` – Number (integer), at least 1.

  The timeout for a task run for this transform in minutes. This is the maximum time that a task run for this transform can consume resources before it is terminated and enters `TIMEOUT` status. The default is 2,880 minutes (48 hours).
- `MaxRetries` – Number (integer).

  The maximum number of times to retry a task for this transform after a task run fails.
- `TransformEncryption` – A TransformEncryption object.

  The encryption-at-rest settings of the transform that apply to accessing user data. Machine learning transforms can access user data encrypted in Amazon S3 using KMS.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`

GetMLTransforms action (Python: get_ml_transforms)

Gets a sortable, filterable list of existing AWS Glue machine learning transforms. Machine learning transforms are a special type of transform that use machine learning to learn the details of the transformation to be performed by learning from examples provided by humans. These transformations are then saved by AWS Glue, and you can retrieve their metadata by calling `GetMLTransforms`.

Request

- `NextToken` – UTF-8 string.

  A paginated token to offset the results.
- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum number of results to return.
- `Filter` – A TransformFilterCriteria object.

  The filter transformation criteria.
- `Sort` – A TransformSortCriteria object.

  The sorting criteria.

Response

- `Transforms` – *Required:* An array of MLTransform objects.

  A list of machine learning transforms.
- `NextToken` – UTF-8 string.

  A pagination token, if more results are available.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`

ListMLTransforms action (Python: list_ml_transforms)

Retrieves a sortable, filterable list of existing AWS Glue machine learning transforms in this AWS account, or the resources with the specified tag. This operation takes the optional `Tags` field, which you can use as a filter of the responses so that tagged resources can be retrieved as a group. If you choose to use tag filtering, only resources with the tags are retrieved.

Request

- `NextToken` – UTF-8 string.

  A continuation token, if this is a continuation request.
- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum size of a list to return.
- `Filter` – A TransformFilterCriteria object.

  A `TransformFilterCriteria` used to filter the machine learning transforms.
- `Sort` – A TransformSortCriteria object.

  A `TransformSortCriteria` used to sort the machine learning transforms.
- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  Specifies to return only these tagged resources.

Response

- `TransformIds` – *Required:* An array of UTF-8 strings.

  The identifiers of all the machine learning transforms in the account, or the machine learning transforms with the specified tags.
- `NextToken` – UTF-8 string.

  A continuation token, if the returned list does not contain the last metric available.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`

StartMLEvaluationTaskRun action (Python: start_ml_evaluation_task_run)

Starts a task to estimate the quality of the transform.

When you provide label sets as examples of truth, AWS Glue machine learning uses some of those examples to learn from them. The rest of the labels are used as a test to estimate quality.

Returns a unique identifier for the run. You can call `GetMLTaskRun` to get more information about the stats of the `EvaluationTaskRun`.

Request

- `TransformId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique identifier of the machine learning transform.

Response

- `TaskRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique identifier associated with this run.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`

- `InternalServiceException`
- `ConcurrentRunsExceededException`
- `MLTransformNotReadyException`

StartMLLabelingSetGenerationTaskRun action (Python: start_ml_labeling_set_generation_task_run)

Starts the active learning workflow for your machine learning transform to improve the transform's quality by generating label sets and adding labels.

When the `StartMLLabelingSetGenerationTaskRun` finishes, AWS Glue will have generated a "labeling set" or a set of questions for humans to answer.

In the case of the `FindMatches` transform, these questions are of the form, "What is the correct way to group these rows together into groups composed entirely of matching records?"

After the labeling process is finished, you can upload your labels with a call to `StartImportLabelsTaskRun`. After `StartImportLabelsTaskRun` finishes, all future runs of the machine learning transform will use the new and improved labels and perform a higher-quality transformation.

Note: The role used to write the generated labeling set to the `OutputS3Path` is the role associated with the Machine Learning Transform, specified in the `CreateMLTransform` API.

Request

- `TransformId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique identifier of the machine learning transform.
- `OutputS3Path` – *Required:* UTF-8 string.

  The Amazon Simple Storage Service (Amazon S3) path where you generate the labeling set.

Response

- `TaskRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The unique run identifier that is associated with this task run.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`
- `ConcurrentRunsExceededException`

GetMLTaskRun action (Python: get_ml_task_run)

Gets details for a specific task run on a machine learning transform. Machine learning task runs are asynchronous tasks that AWS Glue runs on your behalf as part of various machine learning workflows. You can check the stats of any task run by calling `GetMLTaskRun` with the `TaskRunID` and its parent transform's `TransformID`.

Request

- `TransformId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique identifier of the machine learning transform.
- `TaskRunId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique identifier of the task run.

Response

- `TransformId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique identifier of the task run.
- `TaskRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The unique run identifier associated with this run.

- `Status` – UTF-8 string (valid values: `RUNNING` | `FINISHED` | `FAILED` | `PENDING_EXECUTION` | `TIMED_OUT` | `CANCELING` | `CANCELED` | `RECEIVED_BY_TASKRUNNER`).

  The status for this task run.
- `LogGroupName` – UTF-8 string.

  The names of the log groups that are associated with the task run.
- `Properties` – A TaskRunProperties object.

  The list of properties that are associated with the task run.
- `ErrorString` – UTF-8 string.

  The error strings that are associated with the task run.
- `StartedOn` – Timestamp.

  The date and time when this task run started.
- `LastModifiedOn` – Timestamp.

  The date and time when this task run was last modified.
- `CompletedOn` – Timestamp.

  The date and time when this task run was completed.
- `ExecutionTime` – Number (integer).

  The amount of time (in seconds) that the task run consumed resources.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`

GetMLTaskRuns action (Python: get_ml_task_runs)

Gets a list of runs for a machine learning transform. Machine learning task runs are asynchronous tasks that AWS Glue runs on your behalf as part of various machine learning workflows. You can get a sortable, filterable list of machine learning task runs by calling `GetMLTaskRuns` with their parent transform's `TransformID` and other optional parameters as documented in this section.

This operation returns a list of historic runs and must be paginated.

Request

- `TransformId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique identifier of the machine learning transform.
- `NextToken` – UTF-8 string.

  A token for pagination of the results. The default is empty.
- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum number of results to return.
- `Filter` – A TaskRunFilterCriteria object.

  The filter criteria, in the `TaskRunFilterCriteria` structure, for the task run.
- `Sort` – A TaskRunSortCriteria object.

  The sorting criteria, in the `TaskRunSortCriteria` structure, for the task run.

Response

- `TaskRuns` – An array of TaskRun objects.

  A list of task runs that are associated with the transform.
- `NextToken` – UTF-8 string.

  A pagination token, if more results are available.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`

CancelMLTaskRun action (Python: cancel_ml_task_run)

Cancels (stops) a task run. Machine learning task runs are asynchronous tasks that AWS Glue runs on your behalf as part of various machine learning workflows. You can cancel a machine learning task run at any time by calling `CancelMLTaskRun` with a task run's parent transform's `TransformID` and the task run's `TaskRunId`.

Request

- `TransformId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique identifier of the machine learning transform.
- `TaskRunId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A unique identifier for the task run.

Response

- `TransformId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique identifier of the machine learning transform.
- `TaskRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique identifier for the task run.
- `Status` – UTF-8 string (valid values: `RUNNING` | `FINISHED` | `FAILED` | `PENDING_EXECUTION` | `TIMED_OUT` | `CANCELING` | `CANCELED` | `RECEIVED_BY_TASKRUNNER`).

The status for this run.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`

StartExportLabelsTaskRun action (Python: start_export_labels_task_run)

Begins an asynchronous task to export all labeled data for a particular transform. This task is the only label-related API call that is not part of the typical active learning workflow. You typically use `StartExportLabelsTaskRun` when you want to work with all of your existing labels at the same time, such as when you want to remove or change labels that were previously submitted as truth. This API operation accepts the `TransformId` whose labels you want to export and an Amazon Simple Storage Service (Amazon S3) path to export the labels to. The operation returns a `TaskRunId`. You can check on the status of your task run by calling the `GetMLTaskRun` API.

Request

- `TransformId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique identifier of the machine learning transform.
- `OutputS3Path` – *Required:* UTF-8 string.

  The Amazon S3 path where you export the labels.

Response

- `TaskRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique identifier for the task run.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`

StartImportLabelsTaskRun action (Python: start_import_labels_task_run)

Enables you to provide additional labels (examples of truth) to be used to teach the machine learning transform and improve its quality. This API operation is generally used as part of the active learning workflow that starts with the `StartMLLabelingSetGenerationTaskRun` call and that ultimately results in improving the quality of your machine learning transform.

After the `StartMLLabelingSetGenerationTaskRun` finishes, AWS Glue machine learning will have generated a series of questions for humans to answer. (Answering these questions is often called 'labeling' in the machine learning workflows). In the case of the `FindMatches` transform, these questions are of the form, "What is the correct way to group these rows together into groups composed entirely of matching records?" After the labeling process is finished, users upload their answers/labels with a call to `StartImportLabelsTaskRun`. After `StartImportLabelsTaskRun` finishes, all future runs of the machine learning transform use the new and improved labels and perform a higher-quality transformation.

By default, `StartMLLabelingSetGenerationTaskRun` continually learns from and combines all labels that you upload unless you set `Replace` to true. If you set `Replace` to true, `StartImportLabelsTaskRun` deletes and forgets all previously uploaded labels and learns only from the exact set that you upload. Replacing labels can be helpful if you realize that you previously uploaded incorrect labels, and you believe that they are having a negative effect on your transform quality.

You can check on the status of your task run by calling the `GetMLTaskRun` operation.

Request

- `TransformId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The unique identifier of the machine learning transform.

- `InputS3Path` – *Required:* UTF-8 string.

  The Amazon Simple Storage Service (Amazon S3) path from where you import the labels.
- `ReplaceAllLabels` – Boolean.

  Indicates whether to overwrite your existing labels.

Response

- `TaskRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique identifier for the task run.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`
- `ResourceNumberLimitExceededException`
- `InternalServiceException`

Data Quality API

The Data Quality API describes the data quality data types, and includes the API for creating, deleting, or updating data quality rulesets, runs and evaluations.

Data types

- DataSource structure
- DataQualityRulesetListDetails structure
- DataQualityTargetTable structure
- DataQualityRulesetEvaluationRunDescription structure
- DataQualityRulesetEvaluationRunFilter structure

- DataQualityEvaluationRunAdditionalRunOptions structure
- DataQualityRuleRecommendationRunDescription structure
- DataQualityRuleRecommendationRunFilter structure
- DataQualityResult structure
- DataQualityAnalyzerResult structure
- DataQualityObservation structure
- MetricBasedObservation structure
- DataQualityMetricValues structure
- DataQualityRuleResult structure
- DataQualityResultDescription structure
- DataQualityResultFilterCriteria structure
- DataQualityRulesetFilterCriteria structure
- DataQualityAggregatedMetrics structure
- StatisticAnnotation structure
- TimestampedInclusionAnnotation structure
- AnnotationError structure
- DatapointInclusionAnnotation structure
- StatisticSummaryList list
- StatisticSummary structure
- RunIdentifier structure
- StatisticModelResult structure
- DataQualityGlueTable structure

DataSource structure

A data source (an AWS Glue table) for which you want data quality results.

Fields

- `GlueTable` – A GlueTable object.

  An AWS Glue table.
- `DataQualityGlueTable` – A DataQualityGlueTable object.

An AWS Glue table for Data Quality Operations.

DataQualityRulesetListDetails structure

Describes a data quality ruleset returned by `GetDataQualityRuleset`.

Fields

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the data quality ruleset.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the data quality ruleset.
- `CreatedOn` – Timestamp.

  The date and time the data quality ruleset was created.
- `LastModifiedOn` – Timestamp.

  The date and time the data quality ruleset was last modified.
- `TargetTable` – A DataQualityTargetTable object.

  An object representing an AWS Glue table.
- `RecommendationRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  When a ruleset was created from a recommendation run, this run ID is generated to link the two together.
- `RuleCount` – Number (integer).

  The number of rules in the ruleset.

DataQualityTargetTable structure

An object representing an AWS Glue table.

Fields

- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the AWS Glue table.
- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the database where the AWS Glue table exists.
- `CatalogId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The catalog id where the AWS Glue table exists.

DataQualityRulesetEvaluationRunDescription structure

Describes the result of a data quality ruleset evaluation run.

Fields

- `RunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique run identifier associated with this run.
- `Status` – UTF-8 string (valid values: `RUNNING` | `FINISHED` | `FAILED` | `PENDING_EXECUTION` | `TIMED_OUT` | `CANCELING` | `CANCELED` | `RECEIVED_BY_TASKRUNNER`).

  The status for this run.
- `StartedOn` – Timestamp.

  The date and time when the run started.
- `DataSource` – A DataSource object.

  The data source (an AWS Glue table) associated with the run.

DataQualityRulesetEvaluationRunFilter structure

The filter criteria.

Fields

- `DataSource` – *Required:* A DataSource object.

  Filter based on a data source (an AWS Glue table) associated with the run.
- `StartedBefore` – Timestamp.

  Filter results by runs that started before this time.
- `StartedAfter` – Timestamp.

  Filter results by runs that started after this time.

DataQualityEvaluationRunAdditionalRunOptions structure

Additional run options you can specify for an evaluation run.

Fields

- `CloudWatchMetricsEnabled` – Boolean.

  Whether or not to enable CloudWatch metrics.
- `ResultsS3Prefix` – UTF-8 string.

  Prefix for Amazon S3 to store results.
- `CompositeRuleEvaluationMethod` – UTF-8 string (valid values: `COLUMN` | `ROW`).

  Set the evaluation method for composite rules in the ruleset to ROW/COLUMN

DataQualityRuleRecommendationRunDescription structure

Describes the result of a data quality rule recommendation run.

Fields

- `RunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The unique run identifier associated with this run.

- `Status` – UTF-8 string (valid values: `RUNNING` | `FINISHED` | `FAILED` | `PENDING_EXECUTION` | `TIMED_OUT` | `CANCELING` | `CANCELED` | `RECEIVED_BY_TASKRUNNER`).

  The status for this run.
- `StartedOn` – Timestamp.

  The date and time when this run started.
- `DataSource` – A DataSource object.

  The data source (AWS Glue table) associated with the recommendation run.

DataQualityRuleRecommendationRunFilter structure

A filter for listing data quality recommendation runs.

Fields

- `DataSource` – *Required:* A DataSource object.

  Filter based on a specified data source (AWS Glue table).
- `StartedBefore` – Timestamp.

  Filter based on time for results started before provided time.
- `StartedAfter` – Timestamp.

  Filter based on time for results started after provided time.

DataQualityResult structure

Describes a data quality result.

Fields

- `ResultId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A unique result ID for the data quality result.

- `ProfileId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Profile ID for the data quality result.
- `Score` – Number (double), not more than 1.0.

  An aggregate data quality score. Represents the ratio of rules that passed to the total number of rules.
- `DataSource` – A DataSource object.

  The table associated with the data quality result, if any.
- `RulesetName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the ruleset associated with the data quality result.
- `EvaluationContext` – UTF-8 string.

  In the context of a job in AWS Glue Studio, each node in the canvas is typically assigned some sort of name and data quality nodes will have names. In the case of multiple nodes, the `evaluationContext` can differentiate the nodes.
- `StartedOn` – Timestamp.

  The date and time when this data quality run started.
- `CompletedOn` – Timestamp.

  The date and time when this data quality run completed.
- `JobName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The job name associated with the data quality result, if any.
- `JobRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The job run ID associated with the data quality result, if any.
- `RulesetEvaluationRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique run ID for the ruleset evaluation for this data quality result.

- `RuleResults` – An array of DataQualityRuleResult objects, not more than 2000 structures.

  A list of `DataQualityRuleResult` objects representing the results for each rule.
- `AnalyzerResults` – An array of DataQualityAnalyzerResult objects, not more than 2000 structures.

  A list of `DataQualityAnalyzerResult` objects representing the results for each analyzer.
- `Observations` – An array of DataQualityObservation objects, not more than 50 structures.

  A list of `DataQualityObservation` objects representing the observations generated after evaluating the rules and analyzers.
- `AggregatedMetrics` – A DataQualityAggregatedMetrics object.

  A summary of `DataQualityAggregatedMetrics` objects showing the total counts of processed rows and rules, including their pass/fail statistics based on row-level results.

DataQualityAnalyzerResult structure

Describes the result of the evaluation of a data quality analyzer.

Fields

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the data quality analyzer.
- `Description` – UTF-8 string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the data quality analyzer.
- `EvaluationMessage` – UTF-8 string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  An evaluation message.
- `EvaluatedMetrics` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

Each value is a Number (double).

A map of metrics associated with the evaluation of the analyzer.

DataQualityObservation structure

Describes the observation generated after evaluating the rules and analyzers.

Fields

- `Description` – UTF-8 string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the data quality observation.
- `MetricBasedObservation` – A MetricBasedObservation object.

  An object of type `MetricBasedObservation` representing the observation that is based on evaluated data quality metrics.

MetricBasedObservation structure

Describes the metric based observation generated based on evaluated data quality metrics.

Fields

- `MetricName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the data quality metric used for generating the observation.
- `StatisticId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Statistic ID.
- `MetricValues` – A DataQualityMetricValues object.

  An object of type `DataQualityMetricValues` representing the analysis of the data quality metric value.
- `NewRules` – An array of UTF-8 strings.

A list of new data quality rules generated as part of the observation based on the data quality metric value.

DataQualityMetricValues structure

Describes the data quality metric value according to the analysis of historical data.

Fields

- `ActualValue` – Number (double).

  The actual value of the data quality metric.
- `ExpectedValue` – Number (double).

  The expected value of the data quality metric according to the analysis of historical data.
- `LowerLimit` – Number (double).

  The lower limit of the data quality metric value according to the analysis of historical data.
- `UpperLimit` – Number (double).

  The upper limit of the data quality metric value according to the analysis of historical data.

DataQualityRuleResult structure

Describes the result of the evaluation of a data quality rule.

Fields

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the data quality rule.
- `Description` – UTF-8 string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the data quality rule.
- `EvaluationMessage` – UTF-8 string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

An evaluation message.

- `Result` – UTF-8 string (valid values: `PASS` | `FAIL` | `ERROR`).

  A pass or fail status for the rule.
- `EvaluatedMetrics` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a Number (double).

  A map of metrics associated with the evaluation of the rule.
- `EvaluatedRule` – UTF-8 string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  The evaluated rule.
- `RuleMetrics` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a Number (double).

  A map containing metrics associated with the evaluation of the rule based on row-level results.

DataQualityResultDescription structure

Describes a data quality result.

Fields

- `ResultId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique result ID for this data quality result.
- `DataSource` – A DataSource object.

  The table name associated with the data quality result.

- `JobName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The job name associated with the data quality result.
- `JobRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The job run ID associated with the data quality result.
- `StartedOn` – Timestamp.

  The time that the run started for this data quality result.

DataQualityResultFilterCriteria structure

Criteria used to return data quality results.

Fields

- `DataSource` – A DataSource object.

  Filter results by the specified data source. For example, retrieving all results for an AWS Glue table.
- `JobName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Filter results by the specified job name.
- `JobRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Filter results by the specified job run ID.
- `StartedAfter` – Timestamp.

  Filter results by runs that started after this time.
- `StartedBefore` – Timestamp.

  Filter results by runs that started before this time.

DataQualityRulesetFilterCriteria structure

The criteria used to filter data quality rulesets.

Fields

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the ruleset filter criteria.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  The description of the ruleset filter criteria.
- `CreatedBefore` – Timestamp.

  Filter on rulesets created before this date.
- `CreatedAfter` – Timestamp.

  Filter on rulesets created after this date.
- `LastModifiedBefore` – Timestamp.

  Filter on rulesets last modified before this date.
- `LastModifiedAfter` – Timestamp.

  Filter on rulesets last modified after this date.
- `TargetTable` – A DataQualityTargetTable object.

  The name and database name of the target table.

DataQualityAggregatedMetrics structure

A summary of metrics showing the total counts of processed rows and rules, including their pass/fail statistics based on row-level results.

Fields

- `TotalRowsProcessed` – Number (double).

The total number of rows that were processed during the data quality evaluation.

- `TotalRowsPassed` – Number (double).

  The total number of rows that passed all applicable data quality rules.
- `TotalRowsFailed` – Number (double).

  The total number of rows that failed one or more data quality rules.
- `TotalRulesProcessed` – Number (double).

  The total number of data quality rules that were evaluated.
- `TotalRulesPassed` – Number (double).

  The total number of data quality rules that passed their evaluation criteria.
- `TotalRulesFailed` – Number (double).

  The total number of data quality rules that failed their evaluation criteria.

StatisticAnnotation structure

A Statistic Annotation.

Fields

- `ProfileId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Profile ID.
- `StatisticId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Statistic ID.
- `StatisticRecordedOn` – Timestamp.

  The timestamp when the annotated statistic was recorded.
- `InclusionAnnotation` – A TimestampedInclusionAnnotation object.

  The inclusion annotation applied to the statistic.

TimestampedInclusionAnnotation structure

A timestamped inclusion annotation.

Fields

- `Value` – UTF-8 string (valid values: `INCLUDE` | `EXCLUDE`).

  The inclusion annotation value.
- `LastModifiedOn` – Timestamp.

  The timestamp when the inclusion annotation was last modified.

AnnotationError structure

A failed annotation.

Fields

- `ProfileId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Profile ID for the failed annotation.
- `StatisticId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Statistic ID for the failed annotation.
- `FailureReason` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  The reason why the annotation failed.

DatapointInclusionAnnotation structure

An Inclusion Annotation.

Fields

- `ProfileId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The ID of the data quality profile the statistic belongs to.

- `StatisticId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Statistic ID.
- `InclusionAnnotation` – UTF-8 string (valid values: `INCLUDE` | `EXCLUDE`).

  The inclusion annotation value to apply to the statistic.

StatisticSummaryList list

A list of `StatisticSummary`.

An array of StatisticSummary objects.

A list of `StatisticSummary`.

StatisticSummary structure

Summary information about a statistic.

Fields

- `StatisticId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Statistic ID.
- `ProfileId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Profile ID.
- `RunIdentifier` – A RunIdentifier object.

  The Run Identifier
- `StatisticName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Custom string pattern #16.

  The name of the statistic.

- `DoubleValue` – Number (double).

  The value of the statistic.
- `EvaluationLevel` – UTF-8 string (valid values: `Dataset="DATASET"` | `Column="COLUMN"` | `Multicolumn="MULTICOLUMN"`).

  The evaluation level of the statistic. Possible values: `Dataset`, `Column`, `Multicolumn`.
- `ColumnsReferenced` – An array of UTF-8 strings.

  The list of columns referenced by the statistic.
- `ReferencedDatasets` – An array of UTF-8 strings.

  The list of datasets referenced by the statistic.
- `StatisticProperties` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A `StatisticPropertiesMap`, which contains a `NameString` and `DescriptionString`
- `RecordedOn` – Timestamp.

  The timestamp when the statistic was recorded.
- `InclusionAnnotation` – A TimestampedInclusionAnnotation object.

  The inclusion annotation for the statistic.

RunIdentifier structure

A run identifier.

Fields

- `RunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Run ID.

- `JobRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Job Run ID.

StatisticModelResult structure

The statistic model result.

Fields

- `LowerBound` – Number (double).

  The lower bound.
- `UpperBound` – Number (double).

  The upper bound.
- `PredictedValue` – Number (double).

  The predicted value.
- `ActualValue` – Number (double).

  The actual value.
- `Date` – Timestamp.

  The date.
- `InclusionAnnotation` – UTF-8 string (valid values: `INCLUDE` | `EXCLUDE`).

  The inclusion annotation.

DataQualityGlueTable structure

The database and table in the AWS Glue Data Catalog that is used for input or output data for Data Quality Operations.

Fields

- `DatabaseName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

A database name in the AWS Glue Data Catalog.

- `TableName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A table name in the AWS Glue Data Catalog.
- `CatalogId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A unique identifier for the AWS Glue Data Catalog.
- `ConnectionName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the connection to the AWS Glue Data Catalog.
- `AdditionalOptions` – A map array of key-value pairs, not less than 1 or more than 10 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  Additional options for the table. Currently there are two keys supported:
  - `pushDownPredicate`: to filter on partitions without having to list and read all the files in your dataset.
  - `catalogPartitionPredicate`: to use server-side partition pruning using partition indexes in the AWS Glue Data Catalog.
- `PreProcessingQuery` – UTF-8 string, not more than 51200 bytes long, matching the URI address multi-line string pattern.

  SQL Query of SparkSQL format that can be used to pre-process the data for the table in AWS Glue Data Catalog, before running the Data Quality Operation.

Operations

- StartDataQualityRulesetEvaluationRun action (Python: start_data_quality_ruleset_evaluation_run)

- CancelDataQualityRulesetEvaluationRun action (Python: cancel_data_quality_ruleset_evaluation_run)
- GetDataQualityRulesetEvaluationRun action (Python: get_data_quality_ruleset_evaluation_run)
- ListDataQualityRulesetEvaluationRuns action (Python: list_data_quality_ruleset_evaluation_runs)
- StartDataQualityRuleRecommendationRun action (Python: start_data_quality_rule_recommendation_run)
- CancelDataQualityRuleRecommendationRun action (Python: cancel_data_quality_rule_recommendation_run)
- GetDataQualityRuleRecommendationRun action (Python: get_data_quality_rule_recommendation_run)
- ListDataQualityRuleRecommendationRuns action (Python: list_data_quality_rule_recommendation_runs)
- GetDataQualityResult action (Python: get_data_quality_result)
- BatchGetDataQualityResult action (Python: batch_get_data_quality_result)
- ListDataQualityResults action (Python: list_data_quality_results)
- CreateDataQualityRuleset action (Python: create_data_quality_ruleset)
- DeleteDataQualityRuleset action (Python: delete_data_quality_ruleset)
- GetDataQualityRuleset action (Python: get_data_quality_ruleset)
- ListDataQualityRulesets action (Python: list_data_quality_rulesets)
- UpdateDataQualityRuleset action (Python: update_data_quality_ruleset)
- ListDataQualityStatistics action (Python: list_data_quality_statistics)
- TimestampFilter structure
- CreateDataQualityRulesetRequest structure
- GetDataQualityRulesetResponse structure
- GetDataQualityResultResponse structure
- StartDataQualityRuleRecommendationRunRequest structure
- GetDataQualityRuleRecommendationRunResponse structure
- BatchPutDataQualityStatisticAnnotation action (Python: batch_put_data_quality_statistic_annotation)
- GetDataQualityModel action (Python: get_data_quality_model)
- GetDataQualityModelResult action (Python: get_data_quality_model_result)

- ListDataQualityStatisticAnnotations action (Python: list_data_quality_statistic_annotations)
- PutDataQualityProfileAnnotation action (Python: put_data_quality_profile_annotation)

StartDataQualityRulesetEvaluationRun action (Python: start_data_quality_ruleset_evaluation_run)

Once you have a ruleset definition (either recommended or your own), you call this operation to evaluate the ruleset against a data source (AWS Glue table). The evaluation computes results which you can retrieve with the `GetDataQualityResult` API.

Request

- `DataSource` – *Required:* A DataSource object.

  The data source (AWS Glue table) associated with this run.
- `Role` – *Required:* UTF-8 string.

  An IAM role supplied to encrypt the results of the run.
- `NumberOfWorkers` – Number (integer).

  The number of `G.1X` workers to be used in the run. The default is 5.
- `Timeout` – Number (integer), at least 1.

  The timeout for a run in minutes. This is the maximum time that a run can consume resources before it is terminated and enters `TIMEOUT` status. The default is 2,880 minutes (48 hours).
- `ClientToken` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Used for idempotency and is recommended to be set to a random ID (such as a UUID) to avoid creating or starting multiple instances of the same resource.
- `AdditionalRunOptions` – A DataQualityEvaluationRunAdditionalRunOptions object.

  Additional run options you can specify for an evaluation run.
- `RulesetNames` – *Required:* An array of UTF-8 strings, not less than 1 or more than 10 strings.

  A list of ruleset names.
- `AdditionalDataSources` – A map array of key-value pairs.

Each key is a UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

Each value is a A DataSource object.

A map of reference strings to additional data sources you can specify for an evaluation run.

Response

- `RunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique run identifier associated with this run.

Errors

- `InvalidInputException`
- `EntityNotFoundException`
- `OperationTimeoutException`
- `InternalServiceException`
- `ConflictException`

CancelDataQualityRulesetEvaluationRun action (Python: cancel_data_quality_ruleset_evaluation_run)

Cancels a run where a ruleset is being evaluated against a data source.

Request

- `RunId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique run identifier associated with this run.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`

GetDataQualityRulesetEvaluationRun action (Python: get_data_quality_ruleset_evaluation_run)

Retrieves a specific run where a ruleset is evaluated against a data source.

Request

- `RunId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique run identifier associated with this run.

Response

- `RunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique run identifier associated with this run.
- `DataSource` – A DataSource object.

  The data source (an AWS Glue table) associated with this evaluation run.
- `Role` – UTF-8 string.

  An IAM role supplied to encrypt the results of the run.
- `NumberOfWorkers` – Number (integer).

  The number of `G.1X` workers to be used in the run. The default is 5.
- `Timeout` – Number (integer), at least 1.

  The timeout for a run in minutes. This is the maximum time that a run can consume resources before it is terminated and enters `TIMEOUT` status. The default is 2,880 minutes (48 hours).

- `AdditionalRunOptions` – A DataQualityEvaluationRunAdditionalRunOptions object.

  Additional run options you can specify for an evaluation run.
- `Status` – UTF-8 string (valid values: `RUNNING` | `FINISHED` | `FAILED` | `PENDING_EXECUTION` | `TIMED_OUT` | `CANCELING` | `CANCELED` | `RECEIVED_BY_TASKRUNNER`).

  The status for this run.
- `ErrorString` – UTF-8 string.

  The error strings that are associated with the run.
- `StartedOn` – Timestamp.

  The date and time when this run started.
- `LastModifiedOn` – Timestamp.

  A timestamp. The last point in time when this data quality rule recommendation run was modified.
- `CompletedOn` – Timestamp.

  The date and time when this run was completed.
- `ExecutionTime` – Number (integer).

  The amount of time (in seconds) that the run consumed resources.
- `RulesetNames` – An array of UTF-8 strings, not less than 1 or more than 10 strings.

  A list of ruleset names for the run. Currently, this parameter takes only one Ruleset name.
- `ResultIds` – An array of UTF-8 strings, not less than 1 or more than 10 strings.

  A list of result IDs for the data quality results for the run.
- `AdditionalDataSources` – A map array of key-value pairs.

  Each key is a UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Each value is a A DataSource object.

  A map of reference strings to additional data sources you can specify for an evaluation run.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`

ListDataQualityRulesetEvaluationRuns action (Python: list_data_quality_ruleset_evaluation_runs)

Lists all the runs meeting the filter criteria, where a ruleset is evaluated against a data source.

Request

- `Filter` – A DataQualityRulesetEvaluationRunFilter object.

  The filter criteria.
- `NextToken` – UTF-8 string.

  A paginated token to offset the results.
- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum number of results to return.

Response

- `Runs` – An array of DataQualityRulesetEvaluationRunDescription objects.

  A list of `DataQualityRulesetEvaluationRunDescription` objects representing data quality ruleset runs.
- `NextToken` – UTF-8 string.

  A pagination token, if more results are available.

Errors

- `InvalidInputException`

- `OperationTimeoutException`
- `InternalServiceException`

StartDataQualityRuleRecommendationRun action (Python: start_data_quality_rule_recommendation_run)

Starts a recommendation run that is used to generate rules when you don't know what rules to write. AWS Glue Data Quality analyzes the data and comes up with recommendations for a potential ruleset. You can then triage the ruleset and modify the generated ruleset to your liking.

Recommendation runs are automatically deleted after 90 days.

Request

The request of the Data Quality rule recommendation request.

- `DataSource` – *Required:* A DataSource object.

  The data source (AWS Glue table) associated with this run.

- `Role` – *Required:* UTF-8 string.

  An IAM role supplied to encrypt the results of the run.

- `NumberOfWorkers` – Number (integer).

  The number of `G.1X` workers to be used in the run. The default is 5.

- `Timeout` – Number (integer), at least 1.

  The timeout for a run in minutes. This is the maximum time that a run can consume resources before it is terminated and enters `TIMEOUT` status. The default is 2,880 minutes (48 hours).

- `CreatedRulesetName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A name for the ruleset.

- `DataQualitySecurityConfiguration` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the security configuration created with the data quality encryption option.

- `ClientToken` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Used for idempotency and is recommended to be set to a random ID (such as a UUID) to avoid creating or starting multiple instances of the same resource.

Response

- `RunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique run identifier associated with this run.

Errors

- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`
- `ConflictException`

CancelDataQualityRuleRecommendationRun action (Python: cancel_data_quality_rule_recommendation_run)

Cancels the specified recommendation run that was being used to generate rules.

Request

- `RunId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique run identifier associated with this run.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`

GetDataQualityRuleRecommendationRun action (Python: get_data_quality_rule_recommendation_run)

Gets the specified recommendation run that was used to generate rules.

Request

- `RunId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique run identifier associated with this run.

Response

The response for the Data Quality rule recommendation run.

- `RunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique run identifier associated with this run.
- `DataSource` – A DataSource object.

  The data source (an AWS Glue table) associated with this run.
- `Role` – UTF-8 string.

  An IAM role supplied to encrypt the results of the run.
- `NumberOfWorkers` – Number (integer).

  The number of `G.1X` workers to be used in the run. The default is 5.
- `Timeout` – Number (integer), at least 1.

The timeout for a run in minutes. This is the maximum time that a run can consume resources before it is terminated and enters `TIMEOUT` status. The default is 2,880 minutes (48 hours).

- `Status` – UTF-8 string (valid values: `RUNNING` | `FINISHED` | `FAILED` | `PENDING_EXECUTION` | `TIMED_OUT` | `CANCELING` | `CANCELED` | `RECEIVED_BY_TASKRUNNER`).

  The status for this run.
- `ErrorString` – UTF-8 string.

  The error strings that are associated with the run.
- `StartedOn` – Timestamp.

  The date and time when this run started.
- `LastModifiedOn` – Timestamp.

  A timestamp. The last point in time when this data quality rule recommendation run was modified.
- `CompletedOn` – Timestamp.

  The date and time when this run was completed.
- `ExecutionTime` – Number (integer).

  The amount of time (in seconds) that the run consumed resources.
- `RecommendedRuleset` – UTF-8 string, not less than 1 or more than 65536 bytes long.

  When a start rule recommendation run completes, it creates a recommended ruleset (a set of rules). This member has those rules in Data Quality Definition Language (DQDL) format.
- `CreatedRulesetName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the ruleset that was created by the run.
- `DataQualitySecurityConfiguration` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the security configuration created with the data quality encryption option.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`

ListDataQualityRuleRecommendationRuns action (Python: list_data_quality_rule_recommendation_runs)

Lists the recommendation runs meeting the filter criteria.

Request

- `Filter` – A DataQualityRuleRecommendationRunFilter object.

  The filter criteria.
- `NextToken` – UTF-8 string.

  A paginated token to offset the results.
- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum number of results to return.

Response

- `Runs` – An array of DataQualityRuleRecommendationRunDescription objects.

  A list of `DataQualityRuleRecommendationRunDescription` objects.
- `NextToken` – UTF-8 string.

  A pagination token, if more results are available.

Errors

- `InvalidInputException`
- `OperationTimeoutException`

- `InternalServiceException`

GetDataQualityResult action (Python: get_data_quality_result)

Retrieves the result of a data quality rule evaluation.

Request

- `ResultId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A unique result ID for the data quality result.

Response

The response for the data quality result.

- `ResultId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A unique result ID for the data quality result.
- `ProfileId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Profile ID for the data quality result.
- `Score` – Number (double), not more than 1.0.

  An aggregate data quality score. Represents the ratio of rules that passed to the total number of rules.
- `DataSource` – A DataSource object.

  The table associated with the data quality result, if any.
- `RulesetName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the ruleset associated with the data quality result.
- `EvaluationContext` – UTF-8 string.

In the context of a job in AWS Glue Studio, each node in the canvas is typically assigned some sort of name and data quality nodes will have names. In the case of multiple nodes, the `evaluationContext` can differentiate the nodes.

- `StartedOn` – Timestamp.

  The date and time when the run for this data quality result started.
- `CompletedOn` – Timestamp.

  The date and time when the run for this data quality result was completed.
- `JobName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The job name associated with the data quality result, if any.
- `JobRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The job run ID associated with the data quality result, if any.
- `RulesetEvaluationRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique run ID associated with the ruleset evaluation.
- `RuleResults` – An array of DataQualityRuleResult objects, not more than 2000 structures.

  A list of `DataQualityRuleResult` objects representing the results for each rule.
- `AnalyzerResults` – An array of DataQualityAnalyzerResult objects, not more than 2000 structures.

  A list of `DataQualityAnalyzerResult` objects representing the results for each analyzer.
- `Observations` – An array of DataQualityObservation objects, not more than 50 structures.

  A list of `DataQualityObservation` objects representing the observations generated after evaluating the rules and analyzers.
- `AggregatedMetrics` – A DataQualityAggregatedMetrics object.

  A summary of `DataQualityAggregatedMetrics` objects showing the total counts of processed rows and rules, including their pass/fail statistics based on row-level results.

Errors

- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`
- `EntityNotFoundException`

BatchGetDataQualityResult action (Python: batch_get_data_quality_result)

Retrieves a list of data quality results for the specified result IDs.

Request

- `ResultIds` – *Required:* An array of UTF-8 strings, not less than 1 or more than 100 strings.

  A list of unique result IDs for the data quality results.

Response

- `Results` – *Required:* An array of DataQualityResult objects.

  A list of `DataQualityResult` objects representing the data quality results.
- `ResultsNotFound` – An array of UTF-8 strings, not less than 1 or more than 100 strings.

  A list of result IDs for which results were not found.

Errors

- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`

ListDataQualityResults action (Python: list_data_quality_results)

Returns all data quality execution results for your account.

Request

- `Filter` – A DataQualityResultFilterCriteria object.

  The filter criteria.
- `NextToken` – UTF-8 string.

  A paginated token to offset the results.
- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum number of results to return.

Response

- `Results` – *Required:* An array of DataQualityResultDescription objects.

  A list of `DataQualityResultDescription` objects.
- `NextToken` – UTF-8 string.

  A pagination token, if more results are available.

Errors

- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`

CreateDataQualityRuleset action (Python: create_data_quality_ruleset)

Creates a data quality ruleset with DQDL rules applied to a specified AWS Glue table.

You create the ruleset using the Data Quality Definition Language (DQDL). For more information, see the AWS Glue developer guide.

Request

A request to create a data quality ruleset.

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A unique name for the data quality ruleset.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the data quality ruleset.
- `Ruleset` – *Required:* UTF-8 string, not less than 1 or more than 65536 bytes long.

  A Data Quality Definition Language (DQDL) ruleset. For more information, see the AWS Glue developer guide.
- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  A list of tags applied to the data quality ruleset.
- `TargetTable` – A DataQualityTargetTable object.

  A target table associated with the data quality ruleset.
- `RecommendationRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A unique run ID for the recommendation run.
- `DataQualitySecurityConfiguration` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the security configuration created with the data quality encryption option.
- `ClientToken` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Used for idempotency and is recommended to be set to a random ID (such as a UUID) to avoid creating or starting multiple instances of the same resource.

Response

- Name – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A unique name for the data quality ruleset.

Errors

- `InvalidInputException`
- `AlreadyExistsException`
- `OperationTimeoutException`
- `InternalServiceException`
- `ResourceNumberLimitExceededException`

DeleteDataQualityRuleset action (Python: delete_data_quality_ruleset)

Deletes a data quality ruleset.

Request

- Name – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A name for the data quality ruleset.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`

GetDataQualityRuleset action (Python: get_data_quality_ruleset)

Returns an existing ruleset by identifier or name.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the ruleset.

Response

Returns the data quality ruleset response.

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the ruleset.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the ruleset.
- `Ruleset` – UTF-8 string, not less than 1 or more than 65536 bytes long.

  A Data Quality Definition Language (DQDL) ruleset. For more information, see the AWS Glue developer guide.
- `TargetTable` – A DataQualityTargetTable object.

  The name and database name of the target table.
- `CreatedOn` – Timestamp.

  A timestamp. The time and date that this data quality ruleset was created.
- `LastModifiedOn` – Timestamp.

  A timestamp. The last point in time when this data quality ruleset was modified.
- `RecommendationRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

When a ruleset was created from a recommendation run, this run ID is generated to link the two together.

- `DataQualitySecurityConfiguration` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the security configuration created with the data quality encryption option.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`

ListDataQualityRulesets action (Python: list_data_quality_rulesets)

Returns a paginated list of rulesets for the specified list of AWS Glue tables.

Request

- `NextToken` – UTF-8 string.

  A paginated token to offset the results.
- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum number of results to return.
- `Filter` – A DataQualityRulesetFilterCriteria object.

  The filter criteria.
- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  A list of key-value pair tags.

Response

- `Rulesets` – An array of DataQualityRulesetListDetails objects.

  A paginated list of rulesets for the specified list of AWS Glue tables.
- `NextToken` – UTF-8 string.

  A pagination token, if more results are available.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`

UpdateDataQualityRuleset action (Python: update_data_quality_ruleset)

Updates the specified data quality ruleset.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the data quality ruleset.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the ruleset.
- `Ruleset` – UTF-8 string, not less than 1 or more than 65536 bytes long.

  A Data Quality Definition Language (DQDL) ruleset. For more information, see the AWS Glue developer guide.

Response

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the data quality ruleset.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the ruleset.
- `Ruleset` – UTF-8 string, not less than 1 or more than 65536 bytes long.

  A Data Quality Definition Language (DQDL) ruleset. For more information, see the AWS Glue developer guide.

Errors

- `EntityNotFoundException`
- `AlreadyExistsException`
- `IdempotentParameterMismatchException`
- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`
- `ResourceNumberLimitExceededException`

ListDataQualityStatistics action (Python: list_data_quality_statistics)

Retrieves a list of data quality statistics.

Request

- `StatisticId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Statistic ID.
- `ProfileId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

The Profile ID.

- `TimestampFilter` – A TimestampFilter object.

  A timestamp filter.

- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum number of results to return in this request.

- `NextToken` – UTF-8 string.

  A pagination token to request the next page of results.

Response

- `Statistics` – An array of StatisticSummary objects.

  A `StatisticSummaryList`.

- `NextToken` – UTF-8 string.

  A pagination token to request the next page of results.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`

TimestampFilter structure

A timestamp filter.

Fields

- `RecordedBefore` – Timestamp.

  The timestamp before which statistics should be included in the results.

- `RecordedAfter` – Timestamp.

The timestamp after which statistics should be included in the results.

CreateDataQualityRulesetRequest structure

A request to create a data quality ruleset.

Fields

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A unique name for the data quality ruleset.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the data quality ruleset.
- `Ruleset` – *Required:* UTF-8 string, not less than 1 or more than 65536 bytes long.

  A Data Quality Definition Language (DQDL) ruleset. For more information, see the AWS Glue developer guide.
- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  A list of tags applied to the data quality ruleset.
- `TargetTable` – A DataQualityTargetTable object.

  A target table associated with the data quality ruleset.
- `RecommendationRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A unique run ID for the recommendation run.
- `DataQualitySecurityConfiguration` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the security configuration created with the data quality encryption option.

- `ClientToken` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Used for idempotency and is recommended to be set to a random ID (such as a UUID) to avoid creating or starting multiple instances of the same resource.

GetDataQualityRulesetResponse structure

Returns the data quality ruleset response.

Fields

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the ruleset.
- `Description` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A description of the ruleset.
- `Ruleset` – UTF-8 string, not less than 1 or more than 65536 bytes long.

  A Data Quality Definition Language (DQDL) ruleset. For more information, see the AWS Glue developer guide.
- `TargetTable` – A DataQualityTargetTable object.

  The name and database name of the target table.
- `CreatedOn` – Timestamp.

  A timestamp. The time and date that this data quality ruleset was created.
- `LastModifiedOn` – Timestamp.

  A timestamp. The last point in time when this data quality ruleset was modified.
- `RecommendationRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  When a ruleset was created from a recommendation run, this run ID is generated to link the two together.

- `DataQualitySecurityConfiguration` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the security configuration created with the data quality encryption option.

GetDataQualityResultResponse structure

The response for the data quality result.

Fields

- `ResultId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A unique result ID for the data quality result.
- `ProfileId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Profile ID for the data quality result.
- `Score` – Number (double), not more than 1.0.

  An aggregate data quality score. Represents the ratio of rules that passed to the total number of rules.
- `DataSource` – A DataSource object.

  The table associated with the data quality result, if any.
- `RulesetName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the ruleset associated with the data quality result.
- `EvaluationContext` – UTF-8 string.

  In the context of a job in AWS Glue Studio, each node in the canvas is typically assigned some sort of name and data quality nodes will have names. In the case of multiple nodes, the `evaluationContext` can differentiate the nodes.
- `StartedOn` – Timestamp.

  The date and time when the run for this data quality result started.

- `CompletedOn` – Timestamp.

  The date and time when the run for this data quality result was completed.
- `JobName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The job name associated with the data quality result, if any.
- `JobRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The job run ID associated with the data quality result, if any.
- `RulesetEvaluationRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique run ID associated with the ruleset evaluation.
- `RuleResults` – An array of DataQualityRuleResult objects, not more than 2000 structures.

  A list of `DataQualityRuleResult` objects representing the results for each rule.
- `AnalyzerResults` – An array of DataQualityAnalyzerResult objects, not more than 2000 structures.

  A list of `DataQualityAnalyzerResult` objects representing the results for each analyzer.
- `Observations` – An array of DataQualityObservation objects, not more than 50 structures.

  A list of `DataQualityObservation` objects representing the observations generated after evaluating the rules and analyzers.
- `AggregatedMetrics` – A DataQualityAggregatedMetrics object.

  A summary of `DataQualityAggregatedMetrics` objects showing the total counts of processed rows and rules, including their pass/fail statistics based on row-level results.

StartDataQualityRuleRecommendationRunRequest structure

The request of the Data Quality rule recommendation request.

Fields

- `DataSource` – *Required:* A DataSource object.

The data source (AWS Glue table) associated with this run.

- `Role` – *Required:* UTF-8 string.

  An IAM role supplied to encrypt the results of the run.
- `NumberOfWorkers` – Number (integer).

  The number of `G.1X` workers to be used in the run. The default is 5.
- `Timeout` – Number (integer), at least 1.

  The timeout for a run in minutes. This is the maximum time that a run can consume resources before it is terminated and enters `TIMEOUT` status. The default is 2,880 minutes (48 hours).
- `CreatedRulesetName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A name for the ruleset.
- `DataQualitySecurityConfiguration` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the security configuration created with the data quality encryption option.
- `ClientToken` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Used for idempotency and is recommended to be set to a random ID (such as a UUID) to avoid creating or starting multiple instances of the same resource.

GetDataQualityRuleRecommendationRunResponse structure

The response for the Data Quality rule recommendation run.

Fields

- `RunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The unique run identifier associated with this run.
- `DataSource` – A DataSource object.

  The data source (an AWS Glue table) associated with this run.

- `Role` – UTF-8 string.

  An IAM role supplied to encrypt the results of the run.
- `NumberOfWorkers` – Number (integer).

  The number of `G.1X` workers to be used in the run. The default is 5.
- `Timeout` – Number (integer), at least 1.

  The timeout for a run in minutes. This is the maximum time that a run can consume resources before it is terminated and enters `TIMEOUT` status. The default is 2,880 minutes (48 hours).
- `Status` – UTF-8 string (valid values: `RUNNING` | `FINISHED` | `FAILED` | `PENDING_EXECUTION` | `TIMED_OUT` | `CANCELING` | `CANCELED` | `RECEIVED_BY_TASKRUNNER`).

  The status for this run.
- `ErrorString` – UTF-8 string.

  The error strings that are associated with the run.
- `StartedOn` – Timestamp.

  The date and time when this run started.
- `LastModifiedOn` – Timestamp.

  A timestamp. The last point in time when this data quality rule recommendation run was modified.
- `CompletedOn` – Timestamp.

  The date and time when this run was completed.
- `ExecutionTime` – Number (integer).

  The amount of time (in seconds) that the run consumed resources.
- `RecommendedRuleset` – UTF-8 string, not less than 1 or more than 65536 bytes long.

  When a start rule recommendation run completes, it creates a recommended ruleset (a set of rules). This member has those rules in Data Quality Definition Language (DQDL) format.
- `CreatedRulesetName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the ruleset that was created by the run.

- `DataQualitySecurityConfiguration` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the security configuration created with the data quality encryption option.

BatchPutDataQualityStatisticAnnotation action (Python: batch_put_data_quality_statistic_annotation)

Annotate datapoints over time for a specific data quality statistic. The API requires both profileID and statisticID as part of the InclusionAnnotation input. The API only works for a single statisticId across multiple profiles.

Request

- `InclusionAnnotations` – *Required:* An array of DatapointInclusionAnnotation objects.

  A list of `DatapointInclusionAnnotation`'s. The InclusionAnnotations must contain a profileId and statisticId. If there are multiple InclusionAnnotations, the list must refer to a single statisticId across multiple profileIds.
- `ClientToken` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Client Token.

Response

- `FailedInclusionAnnotations` – An array of AnnotationError objects.

  A list of `AnnotationError`'s.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`
- `ResourceNumberLimitExceededException`

GetDataQualityModel action (Python: get_data_quality_model)

Retrieve the training status of the model along with more information (CompletedOn, StartedOn, FailureReason).

Request

- `StatisticId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Statistic ID.
- `ProfileId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Profile ID.

Response

- `Status` – UTF-8 string (valid values: `RUNNING` | `SUCCEEDED` | `FAILED`).

  The training status of the data quality model.
- `StartedOn` – Timestamp.

  The timestamp when the data quality model training started.
- `CompletedOn` – Timestamp.

  The timestamp when the data quality model training completed.
- `FailureReason` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The training failure reason.

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`

GetDataQualityModelResult action (Python: get_data_quality_model_result)

Retrieve a statistic's predictions for a given Profile ID.

Request

- `StatisticId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Statistic ID.
- `ProfileId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Profile ID.

Response

- `CompletedOn` – Timestamp.

  The timestamp when the data quality model training completed.
- `Model` – An array of StatisticModelResult objects.

  A list of `StatisticModelResult`

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`

ListDataQualityStatisticAnnotations action (Python: list_data_quality_statistic_annotations)

Retrieve annotations for a data quality statistic.

Request

- `StatisticId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Statistic ID.
- `ProfileId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The Profile ID.
- `TimestampFilter` – A TimestampFilter object.

  A timestamp filter.
- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum number of results to return in this request.
- `NextToken` – UTF-8 string.

  A pagination token to retrieve the next set of results.

Response

- `Annotations` – An array of StatisticAnnotation objects.

  A list of `StatisticAnnotation` applied to the Statistic
- `NextToken` – UTF-8 string.

  A pagination token to retrieve the next set of results.

Errors

- `InvalidInputException`
- `InternalServiceException`

PutDataQualityProfileAnnotation action (Python: put_data_quality_profile_annotation)

Annotate all datapoints for a Profile.

Request

- `ProfileId` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the data quality monitoring profile to annotate.

- `InclusionAnnotation` – *Required:* UTF-8 string (valid values: `INCLUDE` | `EXCLUDE`).

  The inclusion annotation value to apply to the profile.

Response

- *No Response parameters.*

Errors

- `EntityNotFoundException`
- `InvalidInputException`
- `InternalServiceException`

Sensitive data detection API

The Sensitive data detection API describes the APIs used to detect sensitive data across the columns and rows of your structured data.

Data types

- CustomEntityType structure

CustomEntityType structure

An object representing a custom pattern for detecting sensitive data across the columns and rows of your structured data.

Fields

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A name for the custom pattern that allows it to be retrieved or deleted later. This name must be unique per AWS account.
- `RegexString` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A regular expression string that is used for detecting sensitive data in a custom pattern.
- `ContextWords` – An array of UTF-8 strings, not less than 1 or more than 20 strings.

  A list of context words. If none of these context words are found within the vicinity of the regular expression the data will not be detected as sensitive data.

  If no context words are passed only a regular expression is checked.

Operations

- CreateCustomEntityType action (Python: create_custom_entity_type)
- DeleteCustomEntityType action (Python: delete_custom_entity_type)
- GetCustomEntityType action (Python: get_custom_entity_type)
- BatchGetCustomEntityTypes action (Python: batch_get_custom_entity_types)
- ListCustomEntityTypes action (Python: list_custom_entity_types)

CreateCustomEntityType action (Python: create_custom_entity_type)

Creates a custom pattern that is used to detect sensitive data across the columns and rows of your structured data.

Each custom pattern you create specifies a regular expression and an optional list of context words. If no context words are passed only a regular expression is checked.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A name for the custom pattern that allows it to be retrieved or deleted later. This name must be unique per AWS account.
- `RegexString` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A regular expression string that is used for detecting sensitive data in a custom pattern.
- `ContextWords` – An array of UTF-8 strings, not less than 1 or more than 20 strings.

  A list of context words. If none of these context words are found within the vicinity of the regular expression the data will not be detected as sensitive data.

  If no context words are passed only a regular expression is checked.
- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  A list of tags applied to the custom entity type.

Response

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the custom pattern you created.

Errors

- `AccessDeniedException`
- `AlreadyExistsException`

- `IdempotentParameterMismatchException`
- `InternalServiceException`
- `InvalidInputException`
- `OperationTimeoutException`
- `ResourceNumberLimitExceededException`

DeleteCustomEntityType action (Python: delete_custom_entity_type)

Deletes a custom pattern by specifying its name.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the custom pattern that you want to delete.

Response

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the custom pattern you deleted.

Errors

- `EntityNotFoundException`
- `AccessDeniedException`
- `InternalServiceException`
- `InvalidInputException`
- `OperationTimeoutException`

GetCustomEntityType action (Python: get_custom_entity_type)

Retrieves the details of a custom pattern by specifying its name.

Request

- `Name` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the custom pattern that you want to retrieve.

Response

- `Name` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the custom pattern that you retrieved.
- `RegexString` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  A regular expression string that is used for detecting sensitive data in a custom pattern.
- `ContextWords` – An array of UTF-8 strings, not less than 1 or more than 20 strings.

  A list of context words if specified when you created the custom pattern. If none of these context words are found within the vicinity of the regular expression the data will not be detected as sensitive data.

Errors

- `EntityNotFoundException`
- `AccessDeniedException`
- `InternalServiceException`
- `InvalidInputException`
- `OperationTimeoutException`

BatchGetCustomEntityTypes action (Python: batch_get_custom_entity_types)

Retrieves the details for the custom patterns specified by a list of names.

Request

- `Names` – *Required:* An array of UTF-8 strings, not less than 1 or more than 50 strings.

  A list of names of the custom patterns that you want to retrieve.

Response

- `CustomEntityTypes` – An array of CustomEntityType objects.

  A list of `CustomEntityType` objects representing the custom patterns that have been created.
- `CustomEntityTypesNotFound` – An array of UTF-8 strings, not less than 1 or more than 50 strings.

  A list of the names of custom patterns that were not found.

Errors

- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`

ListCustomEntityTypes action (Python: list_custom_entity_types)

Lists all the custom patterns that have been created.

Request

- `NextToken` – UTF-8 string.

  A paginated token to offset the results.
- `MaxResults` – Number (integer), not less than 1 or more than 1000.

  The maximum number of results to return.
- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

A list of key-value pair tags.

Response

- `CustomEntityTypes` – An array of CustomEntityType objects.

  A list of `CustomEntityType` objects representing custom patterns.
- `NextToken` – UTF-8 string.

  A pagination token, if more results are available.

Errors

- `InvalidInputException`
- `OperationTimeoutException`
- `InternalServiceException`

Tagging APIs in AWS Glue

Data types

- Tag structure

Tag structure

The `Tag` object represents a label that you can assign to an AWS resource. Each tag consists of a key and an optional value, both of which you define.

For more information about tags, and controlling access to resources in AWS Glue, see AWS Tags in AWS Glue and Specifying AWS Glue Resource ARNs in the developer guide.

Fields

- `key` – UTF-8 string, not less than 1 or more than 128 bytes long.

  The tag key. The key is required when you create a tag on an object. The key is case-sensitive, and must not contain the prefix aws.

- `value` – UTF-8 string, not more than 256 bytes long.

  The tag value. The value is optional when you create a tag on an object. The value is case-sensitive, and must not contain the prefix aws.

Operations

- TagResource action (Python: tag_resource)
- UntagResource action (Python: untag_resource)
- GetTags action (Python: get_tags)

TagResource action (Python: tag_resource)

Adds tags to a resource. A tag is a label you can assign to an AWS resource. In AWS Glue, you can tag only certain resources. For information about what resources you can tag, see AWS Tags in AWS Glue.

In addition to the tagging permissions to call tag related APIs, you also need the `glue:GetConnection` permission to call tagging APIs on connections, and the `glue:GetDatabase` permission to call tagging APIs on databases.

Request

- `ResourceArn` – *Required:* UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

  The ARN of the AWS Glue resource to which to add the tags. For more information about AWS Glue resource ARNs, see the AWS Glue ARN string pattern.
- `TagsToAdd` – *Required:* A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  Tags to add to this resource.

Response

- *No Response parameters.*

Errors

- `ResourceNotFoundException`

UntagResource action (Python: untag_resource)

Removes the specified tags from an integration resource.

Request

- `ResourceArn` – *Required:* UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

  The Amazon Resource Name (ARN) for the integration resource.
- `TagsToRemove` – *Required:* An array of UTF-8 strings, not more than 50 strings.

  A list of metadata tags to be removed from the resource.

Response

- *No Response parameters.*

Errors

- `ResourceNotFoundException`

GetTags action (Python: get_tags)

Retrieves a list of tags associated with a resource.

Request

- `ResourceArn` – *Required:* UTF-8 string, not less than 1 or more than 10240 bytes long, matching the Custom string pattern #49.

The Amazon Resource Name (ARN) of the resource for which to retrieve tags.

Response

- `Tags` – A map array of key-value pairs, not more than 50 pairs.

  Each key is a UTF-8 string, not less than 1 or more than 128 bytes long.

  Each value is a UTF-8 string, not more than 256 bytes long.

  The requested tags.

Errors

- `InvalidInputException`
- `InternalServiceException`
- `OperationTimeoutException`
- `EntityNotFoundException`

Common data types

The Common data types describes miscellaneous common data types in AWS Glue.

Tag structure

The `Tag` object represents a label that you can assign to an AWS resource. Each tag consists of a key and an optional value, both of which you define.

For more information about tags, and controlling access to resources in AWS Glue, see AWS Tags in AWS Glue and Specifying AWS Glue Resource ARNs in the developer guide.

Fields

- `key` – UTF-8 string, not less than 1 or more than 128 bytes long.

  The tag key. The key is required when you create a tag on an object. The key is case-sensitive, and must not contain the prefix aws.
- `value` – UTF-8 string, not more than 256 bytes long.

The tag value. The value is optional when you create a tag on an object. The value is case-sensitive, and must not contain the prefix aws.

DecimalNumber structure

Contains a numeric value in decimal format.

Fields

- `UnscaledValue` – *Required:* Blob.

  The unscaled numeric value.
- `Scale` – *Required:* Number (integer).

  The scale that determines where the decimal point falls in the unscaled value.

ErrorDetail structure

Contains details about an error.

Fields

- `ErrorCode` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The code associated with this error.
- `ErrorMessage` – Description string, not more than 2048 bytes long, matching the URI address multi-line string pattern.

  A message describing the error.

PropertyPredicate structure

Defines a property predicate.

Fields

- `Key` – Value string, not less than 1 or more than 1024 bytes long.

The key of the property.

- `Value` – Value string, not less than 1 or more than 1024 bytes long.

  The value of the property.
- `Comparator` – UTF-8 string (valid values: `EQUALS` | `GREATER_THAN` | `LESS_THAN` | `GREATER_THAN_EQUALS` | `LESS_THAN_EQUALS`).

  The comparator used to compare this property to others.

ResourceUri structure

The URIs for function resources.

Fields

- `ResourceType` – UTF-8 string (valid values: `JAR` | `FILE` | `ARCHIVE`).

  The type of the resource.
- `Uri` – Uniform resource identifier (uri), not less than 1 or more than 1024 bytes long, matching the URI address multi-line string pattern.

  The URI for accessing the resource.

ColumnStatistics structure

Represents the generated column-level statistics for a table or partition.

Fields

- `ColumnName` – *Required:* UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  Name of column which statistics belong to.
- `ColumnType` – *Required:* Type name, not more than 20000 bytes long, matching the Single-line string pattern.

  The data type of the column.
- `AnalyzedTime` – *Required:* Timestamp.

The timestamp of when column statistics were generated.

- `StatisticsData` – *Required:* A ColumnStatisticsData object.

  A `ColumnStatisticData` object that contains the statistics data values.

ColumnStatisticsError structure

Encapsulates a `ColumnStatistics` object that failed and the reason for failure.

Fields

- `ColumnStatistics` – A ColumnStatistics object.

  The `ColumnStatistics` of the column.
- `Error` – An ErrorDetail object.

  An error message with the reason for the failure of an operation.

ColumnError structure

Encapsulates a column name that failed and the reason for failure.

Fields

- `ColumnName` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The name of the column that failed.
- `Error` – An ErrorDetail object.

  An error message with the reason for the failure of an operation.

ColumnStatisticsData structure

Contains the individual types of column statistics data. Only one data object should be set and indicated by the `Type` attribute.

Fields

- `Type` – *Required:* UTF-8 string (valid values: `BOOLEAN` | `DATE` | `DECIMAL` | `DOUBLE` | `LONG` | `STRING` | `BINARY`).

  The type of column statistics data.
- `BooleanColumnStatisticsData` – A BooleanColumnStatisticsData object.

  Boolean column statistics data.
- `DateColumnStatisticsData` – A DateColumnStatisticsData object.

  Date column statistics data.
- `DecimalColumnStatisticsData` – A DecimalColumnStatisticsData object.

  Decimal column statistics data. UnscaledValues within are Base64-encoded binary objects storing big-endian, two's complement representations of the decimal's unscaled value.
- `DoubleColumnStatisticsData` – A DoubleColumnStatisticsData object.

  Double column statistics data.
- `LongColumnStatisticsData` – A LongColumnStatisticsData object.

  Long column statistics data.
- `StringColumnStatisticsData` – A StringColumnStatisticsData object.

  String column statistics data.
- `BinaryColumnStatisticsData` – A BinaryColumnStatisticsData object.

  Binary column statistics data.

BooleanColumnStatisticsData structure

Defines column statistics supported for Boolean data columns.

Fields

- `NumberOfTrues` – *Required:* Number (long), not more than None.

  The number of true values in the column.
- `NumberOfFalses` – *Required:* Number (long), not more than None.

The number of false values in the column.

- `NumberOfNulls` – *Required:* Number (long), not more than None.

  The number of null values in the column.

DateColumnStatisticsData structure

Defines column statistics supported for timestamp data columns.

Fields

- `MinimumValue` – Timestamp.

  The lowest value in the column.
- `MaximumValue` – Timestamp.

  The highest value in the column.
- `NumberOfNulls` – *Required:* Number (long), not more than None.

  The number of null values in the column.
- `NumberOfDistinctValues` – *Required:* Number (long), not more than None.

  The number of distinct values in a column.

DecimalColumnStatisticsData structure

Defines column statistics supported for fixed-point number data columns.

Fields

- `MinimumValue` – A DecimalNumber object.

  The lowest value in the column.
- `MaximumValue` – A DecimalNumber object.

  The highest value in the column.
- `NumberOfNulls` – *Required:* Number (long), not more than None.

The number of null values in the column.

- `NumberOfDistinctValues` – *Required:* Number (long), not more than None.

  The number of distinct values in a column.

DoubleColumnStatisticsData structure

Defines column statistics supported for floating-point number data columns.

Fields

- `MinimumValue` – Number (double).

  The lowest value in the column.
- `MaximumValue` – Number (double).

  The highest value in the column.
- `NumberOfNulls` – *Required:* Number (long), not more than None.

  The number of null values in the column.
- `NumberOfDistinctValues` – *Required:* Number (long), not more than None.

  The number of distinct values in a column.

LongColumnStatisticsData structure

Defines column statistics supported for integer data columns.

Fields

- `MinimumValue` – Number (long).

  The lowest value in the column.
- `MaximumValue` – Number (long).

  The highest value in the column.
- `NumberOfNulls` – *Required:* Number (long), not more than None.

The number of null values in the column.

- `NumberOfDistinctValues` – *Required:* Number (long), not more than None.

  The number of distinct values in a column.

StringColumnStatisticsData structure

Defines column statistics supported for character sequence data values.

Fields

- `MaximumLength` – *Required:* Number (long), not more than None.

  The size of the longest string in the column.
- `AverageLength` – *Required:* Number (double), not more than None.

  The average string length in the column.
- `NumberOfNulls` – *Required:* Number (long), not more than None.

  The number of null values in the column.
- `NumberOfDistinctValues` – *Required:* Number (long), not more than None.

  The number of distinct values in a column.

BinaryColumnStatisticsData structure

Defines column statistics supported for bit sequence data values.

Fields

- `MaximumLength` – *Required:* Number (long), not more than None.

  The size of the longest bit sequence in the column.
- `AverageLength` – *Required:* Number (double), not more than None.

  The average bit sequence length in the column.
- `NumberOfNulls` – *Required:* Number (long), not more than None.

The number of null values in the column.

String patterns

The API uses the following regular expressions to define what is valid content for various string parameters and members:

- Single-line string pattern – "`[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\t]*`"
- URI address multi-line string pattern – "`[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`"
- A Logstash Grok string pattern – "`[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\t]*`"
- Identifier string pattern – "`[A-Za-z_][A-Za-z0-9_]*`"
- AWS IAM ARN string pattern – "`arn:aws:iam::\d{12}:role/.*`"
- Version string pattern – "`^[a-zA-Z0-9-_]+$`"
- Log group string pattern – "`[\.\-_/#A-Za-z0-9]+`"
- Log-stream string pattern – "`[^:*]*`"
- Custom string pattern #10 – "`[a-zA-Z0-9-_]+`"
- Custom string pattern #11 – "`[-a-zA-Z0-9+=/:_]*`"
- Custom string pattern #12 – "`[\S\s]*`"
- Custom string pattern #13 – "`.*\S.*`"
- Custom string pattern #14 – "`[a-zA-Z0-9-=._/@]+`"
- Custom string pattern #15 – "`[1-9][0-9]*|[1-9][0-9]*-[1-9][0-9]*`"
- Custom string pattern #16 – "`[A-Z][A-Za-z\.]+`"
- Custom string pattern #17 – "`[\S]*`"
- Custom string pattern #18 – "`[\w]*`"
- Custom string pattern #19 – "`arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`"
- Custom string pattern #20 – "`subnet-[a-z0-9]+`"
- Custom string pattern #21 – "`\d{12}`"

- Custom string pattern #22 – `"([a-z]+)-([a-z]+-)?([a-z]+)-[0-9]+[a-z]+"`
- Custom string pattern #23 – `"[a-zA-Z0-9.-]*"`
- Custom string pattern #24 – `"arn:aws[a-z0-9\-]*:lambda:[a-z0-9\-]+:\d{12}:function:([\w\-]{1,64})"`
- Custom string pattern #25 – `"^(?!(.*[.\/\\]|aws:)).*$"`
- Custom string pattern #26 – `"[^\r\n]"`
- Custom string pattern #27 – `"^\w+\.\w+\.\w+$"`
- Custom string pattern #28 – `"^\w+\.\w+$"`
- Custom string pattern #29 – `"^$|arn:aws[a-z0-9-]*:kms:.*"`
- Custom string pattern #30 – `"arn:aws[^:]*:iam::[0-9]*:role/.+"`
- Custom string pattern #31 – `"[\.\-_A-Za-z0-9]+"`
- Custom string pattern #32 – `"^s3://([^/]+)/([^/]+/)*([^/]+)$"`
- Custom string pattern #33 – `".*"`
- Custom string pattern #34 – `"^(Sun|Mon|Tue|Wed|Thu|Fri|Sat):([01]?[0-9]|2[0-3])$"`
- Custom string pattern #35 – `"[a-zA-Z0-9_.-]+"`
- Custom string pattern #36 – `"^arn:aws(-(cn|us-gov|iso(-[bef])?))?:secretsmanager:.*$"`
- Custom string pattern #37 – `"\S+"`
- Custom string pattern #38 – `"^[\x20-\x7E]*$"`
- Custom string pattern #39 – `"^([a-zA-Z0-9_=]+)\.([a-zA-Z0-9_=]+)\.([a-zA-Z0-9_\-\+\/=]*)"`
- Custom string pattern #40 – `"^(https?)://[-a-zA-Z0-9+&@#/%?=~_|!:,.;]*[-a-zA-Z0-9+&@#/%=~_|]"`
- Custom string pattern #41 – `"^(https?):\/\/[^\s/$.?#].[^\s]*$"`
- Custom string pattern #42 – `"^subnet-[a-z0-9]+$"`
- Custom string pattern #43 – `"[\p{L}\p{N}\p{P}]*"`
- Custom string pattern #44 – `"[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}"`
- Custom string pattern #45 – `"[a-zA-Z0-9-_$#.]+"`
- Custom string pattern #46 – `"^\d{12}$"`

- Custom string pattern #47 – `"^(\w+\.)+\w+$"`
- Custom string pattern #48 – `"^([2-3]|3[.]9)$"`
- Custom string pattern #49 – `"arn:aws(-(cn|us-gov|iso(-[bef])?))?:glue:.*"`
- Custom string pattern #50 – `"(^arn:aws(-(cn|us-gov|iso(-[bef])?))?:iam::\w{12}:root)"`
- Custom string pattern #51 – `"^arn:aws(-(cn|us-gov|iso(-[bef])?))?:iam::[0-9]{12}:role/.+"`
- Custom string pattern #52 – `"[\s\S]*"`
- Custom string pattern #53 – `"([\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF]|[^\S\r\n"'= ;])*"`
- Custom string pattern #54 – `"^[A-Z\_]+$"`
- Custom string pattern #55 – `"^[A-Za-z0-9]+$"`
- Custom string pattern #56 – `"[*A-Za-z0-9_-]*"`
- Custom string pattern #57 – `"([\u0020-\u007E\r\s\n])*"`
- Custom string pattern #58 – `"[A-Za-z0-9_-]*"`
- Custom string pattern #59 – `"([\u0009\u000B\u000C\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF])*"`
- Custom string pattern #60 – `"([\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\s])*"`
- Custom string pattern #61 – `"([^\r\n])*"`

Exceptions

This section describes AWS Glue exceptions that you can use to find the source of problems and fix them. For more information on HTTP error codes and strings for exceptions related to machine learning, see the section called "AWS Glue machine learning exceptions".

AccessDeniedException structure

Access to a resource was denied.

Fields

- `Message` – UTF-8 string.

A message describing the problem.

AlreadyExistsException structure

A resource to be created or added already exists.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

ConcurrentModificationException structure

Two processes are trying to modify a resource simultaneously.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

ConcurrentRunsExceededException structure

Too many jobs are being run concurrently.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

CrawlerNotRunningException structure

The specified crawler is not running.

Fields

- `Message` – UTF-8 string.

A message describing the problem.

CrawlerRunningException structure

The operation cannot be performed because the crawler is already running.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

CrawlerStoppingException structure

The specified crawler is stopping.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

EntityNotFoundException structure

A specified entity does not exist

Fields

- `Message` – UTF-8 string.

  A message describing the problem.
- `FromFederationSource` – Boolean.

  Indicates whether or not the exception relates to a federated source.

FederationSourceException structure

A federation source failed.

Fields

- `FederationSourceErrorCode` – UTF-8 string (valid values: `AccessDeniedException` | `EntityNotFoundException` | `InvalidCredentialsException` | `InvalidInputException` | `InvalidResponseException` | `OperationTimeoutException` | `OperationNotSupportedException` | `InternalServiceException` | `PartialFailureException` | `ThrottlingException`).

  The error code of the problem.
- `Message` – UTF-8 string.

  The message describing the problem.

FederationSourceRetryableException structure

A federation source failed, but the operation may be retried.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

GlueEncryptionException structure

An encryption operation failed.

Fields

- `Message` – UTF-8 string.

  The message describing the problem.

IdempotentParameterMismatchException structure

The same unique identifier was associated with two different records.

Fields

- `Message` – UTF-8 string.

A message describing the problem.

IllegalWorkflowStateException structure

The workflow is in an invalid state to perform a requested operation.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

InternalServiceException structure

An internal service error occurred.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

InvalidExecutionEngineException structure

An unknown or invalid execution engine was specified.

Fields

- `message` – UTF-8 string.

  A message describing the problem.

InvalidInputException structure

The input provided was not valid.

Fields

- `Message` – UTF-8 string.

A message describing the problem.

- `FromFederationSource` – Boolean.

  Indicates whether or not the exception relates to a federated source.

InvalidStateException structure

An error that indicates your data is in an invalid state.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

InvalidTaskStatusTransitionException structure

Proper transition from one task to the next failed.

Fields

- `message` – UTF-8 string.

  A message describing the problem.

JobDefinitionErrorException structure

A job definition is not valid.

Fields

- `message` – UTF-8 string.

  A message describing the problem.

JobRunInTerminalStateException structure

The terminal state of a job run signals a failure.

Fields

- `message` – UTF-8 string.

  A message describing the problem.

JobRunInvalidStateTransitionException structure

A job run encountered an invalid transition from source state to target state.

Fields

- `jobRunId` – UTF-8 string, not less than 1 or more than 255 bytes long, matching the Single-line string pattern.

  The ID of the job run in question.
- `message` – UTF-8 string.

  A message describing the problem.
- `sourceState` – UTF-8 string (valid values: `STARTING` | `RUNNING` | `STOPPING` | `STOPPED` | `SUCCEEDED` | `FAILED` | `TIMEOUT` | `ERROR` | `WAITING` | `EXPIRED`).

  The source state.
- `targetState` – UTF-8 string (valid values: `STARTING` | `RUNNING` | `STOPPING` | `STOPPED` | `SUCCEEDED` | `FAILED` | `TIMEOUT` | `ERROR` | `WAITING` | `EXPIRED`).

  The target state.

JobRunNotInTerminalStateException structure

A job run is not in a terminal state.

Fields

- `message` – UTF-8 string.

  A message describing the problem.

LateRunnerException structure

A job runner is late.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

NoScheduleException structure

There is no applicable schedule.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

OperationTimeoutException structure

The operation timed out.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

ResourceNotReadyException structure

A resource was not ready for a transaction.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

ResourceNumberLimitExceededException structure

A resource numerical limit was exceeded.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

SchedulerNotRunningException structure

The specified scheduler is not running.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

SchedulerRunningException structure

The specified scheduler is already running.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

SchedulerTransitioningException structure

The specified scheduler is transitioning.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

UnrecognizedRunnerException structure

The job runner was not recognized.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

ValidationException structure

A value could not be validated.

Fields

- `Message` – UTF-8 string.

  A message describing the problem.

VersionMismatchException structure

There was a version conflict.

Fields

- `message` – UTF-8 string.

  A message describing the problem.

AWS Glue API code examples using AWS SDKs

The following code examples show how to use AWS Glue with an AWS software development kit (SDK).

Basics are code examples that show you how to perform the essential operations within a service.

Actions are code excerpts from larger programs and must be run in context. While actions show you how to call individual service functions, you can see actions in context in their related scenarios.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Get started

Hello AWS Glue

The following code examples show how to get started using AWS Glue.

.NET

SDK for .NET

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
namespace GlueActions;

public class HelloGlue
{
    private static ILogger logger = null!;

    static async Task Main(string[] args)
    {
        // Set up dependency injection for AWS Glue.
        using var host = Host.CreateDefaultBuilder(args)
```

```
            .ConfigureLogging(logging =>
                logging.AddFilter("System", LogLevel.Debug)
                    .AddFilter<DebugLoggerProvider>("Microsoft",
 LogLevel.Information)
                    .AddFilter<ConsoleLoggerProvider>("Microsoft",
 LogLevel.Trace))
            .ConfigureServices((_, services) =>
                services.AddAWSService<IAmazonGlue>()
                .AddTransient<GlueWrapper>()
            )
            .Build();

        logger = LoggerFactory.Create(builder => { builder.AddConsole(); })
            .CreateLogger<HelloGlue>();
        var glueClient = host.Services.GetRequiredService<IAmazonGlue>();

        var request = new ListJobsRequest();

        var jobNames = new List<string>();

        do
        {
            var response = await glueClient.ListJobsAsync(request);
            jobNames.AddRange(response.JobNames);
            request.NextToken = response.NextToken;
        }
        while (request.NextToken is not null);

        Console.Clear();
        Console.WriteLine("Hello, Glue. Let's list your existing Glue Jobs:");
        if (jobNames.Count == 0)
        {
            Console.WriteLine("You don't have any AWS Glue jobs.");
        }
        else
        {
            jobNames.ForEach(Console.WriteLine);
        }
    }
}
```

- For API details, see ListJobs in *AWS SDK for .NET API Reference*.

C++

SDK for C++

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the [AWS Code Examples Repository](https://github.com/awsdocs/aws-doc-sdk-examples).

Code for the CMakeLists.txt CMake file.

```
# Set the minimum required version of CMake for this project.
cmake_minimum_required(VERSION 3.13)

# Set the AWS service components used by this project.
set(SERVICE_COMPONENTS glue)

# Set this project's name.
project("hello_glue")

# Set the C++ standard to use to build this target.
# At least C++ 11 is required for the AWS SDK for C++.
set(CMAKE_CXX_STANDARD 11)

# Use the MSVC variable to determine if this is a Windows build.
set(WINDOWS_BUILD ${MSVC})

if (WINDOWS_BUILD) # Set the location where CMake can find the installed
 libraries for the AWS SDK.
    string(REPLACE ";" "/aws-cpp-sdk-all;" SYSTEM_MODULE_PATH
 "${CMAKE_SYSTEM_PREFIX_PATH}/aws-cpp-sdk-all")
    list(APPEND CMAKE_PREFIX_PATH ${SYSTEM_MODULE_PATH})
endif ()

# Find the AWS SDK for C++ package.
find_package(AWSSDK REQUIRED COMPONENTS ${SERVICE_COMPONENTS})

if (WINDOWS_BUILD AND AWSSDK_INSTALL_AS_SHARED_LIBS)
     # Copy relevant AWS SDK for C++ libraries into the current binary directory
 for running and debugging.
```

```
    # set(BIN_SUB_DIR "/Debug") # if you are building from the command line you
 may need to uncomment this
                                    # and set the proper subdirectory to the
 executables' location.

    AWSSDK_CPY_DYN_LIBS(SERVICE_COMPONENTS ""
 ${CMAKE_CURRENT_BINARY_DIR}${BIN_SUB_DIR})
endif ()

add_executable(${PROJECT_NAME}
        hello_glue.cpp)

target_link_libraries(${PROJECT_NAME}
        ${AWSSDK_LINK_LIBRARIES})
```

Code for the hello_glue.cpp source file.

```
#include <aws/core/Aws.h>
#include <aws/glue/GlueClient.h>
#include <aws/glue/model/ListJobsRequest.h>
#include <iostream>

/*
 *  A "Hello Glue" starter application which initializes an AWS Glue client and
 lists the
 *  AWS Glue job definitions.
 *
 *  main function
 *
 *  Usage: 'hello_glue'
 *
 */

int main(int argc, char **argv) {
    Aws::SDKOptions options;
    // Optionally change the log level for debugging.
//   options.loggingOptions.logLevel = Utils::Logging::LogLevel::Debug;
    Aws::InitAPI(options); // Should only be called once.
    int result = 0;
    {
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region (overrides config file).
        // clientConfig.region = "us-east-1";

        Aws::Glue::GlueClient glueClient(clientConfig);

        std::vector<Aws::String> jobs;

        Aws::String nextToken;  // Used for pagination.
        do {
            Aws::Glue::Model::ListJobsRequest listJobsRequest;
            if (!nextToken.empty()) {
                listJobsRequest.SetNextToken(nextToken);
            }

            Aws::Glue::Model::ListJobsOutcome listRunsOutcome =
 glueClient.ListJobs(
                    listJobsRequest);

            if (listRunsOutcome.IsSuccess()) {
                const std::vector<Aws::String> &jobNames =
 listRunsOutcome.GetResult().GetJobNames();
                jobs.insert(jobs.end(), jobNames.begin(), jobNames.end());

                nextToken = listRunsOutcome.GetResult().GetNextToken();
            } else {
                std::cerr << "Error listing jobs. "
                          << listRunsOutcome.GetError().GetMessage()
                          << std::endl;
                result = 1;
                break;
            }
        } while (!nextToken.empty());

        std::cout << "Your account has " << jobs.size() << " jobs."
                  << std::endl;
        for (size_t i = 0; i < jobs.size(); ++i) {
            std::cout << "   " << i + 1 << ". " << jobs[i] << std::endl;
        }
    }
    Aws::ShutdownAPI(options); // Should only be called once.
    return result;
}
```

- For API details, see ListJobs in *AWS SDK for C++ API Reference*.

Java

SDK for Java 2.x

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
package com.example.glue;

import software.amazon.awssdk.regions.Region;
import software.amazon.awssdk.services.glue.GlueClient;
import software.amazon.awssdk.services.glue.model.ListJobsRequest;
import software.amazon.awssdk.services.glue.model.ListJobsResponse;
import java.util.List;

public class HelloGlue {
    public static void main(String[] args) {
        GlueClient glueClient = GlueClient.builder()
                .region(Region.US_EAST_1)
                .build();

        listJobs(glueClient);
    }

    public static void listJobs(GlueClient glueClient) {
        ListJobsRequest request = ListJobsRequest.builder()
                .maxResults(10)
                .build();
        ListJobsResponse response = glueClient.listJobs(request);
        List<String> jobList = response.jobNames();
        jobList.forEach(job -> {
            System.out.println("Job Name: " + job);
        });
    }
}
```

- For API details, see ListJobs in *AWS SDK for Java 2.x API Reference*.

JavaScript

SDK for JavaScript (v3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
import { ListJobsCommand, GlueClient } from "@aws-sdk/client-glue";

const client = new GlueClient({});

export const main = async () => {
  const command = new ListJobsCommand({});

  const { JobNames } = await client.send(command);
  const formattedJobNames = JobNames.join("\n");
  console.log("Job names: ");
  console.log(formattedJobNames);
  return JobNames;
};
```

- For API details, see ListJobs in *AWS SDK for JavaScript API Reference*.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```

import boto3
from botocore.exceptions import ClientError
```

```
def hello_glue():
    """
    Lists the job definitions in your AWS Glue account, using the AWS SDK for
 Python (Boto3).
    """
    try:
        # Create the Glue client
        glue = boto3.client("glue")

        # List the jobs, limiting the results to 10 per page
        paginator = glue.get_paginator("get_jobs")
        response_iterator = paginator.paginate(
            PaginationConfig={"MaxItems": 10, "PageSize": 10}
        )

        # Print the job names
        print("Here are the jobs in your account:")
        for page in response_iterator:
            for job in page["Jobs"]:
                print(f"\t{job['Name']}")

    except ClientError as e:
        print(f"Error: {e}")


if __name__ == "__main__":
    hello_glue()
```

- For API details, see ListJobs in *AWS SDK for Python (Boto3) API Reference*.

Ruby

SDK for Ruby

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
require 'aws-sdk-glue'
require 'logger'

# GlueManager is a class responsible for managing AWS Glue operations
# such as listing all Glue jobs in the current AWS account.
class GlueManager
  def initialize(client)
    @client = client
    @logger = Logger.new($stdout)
  end

  # Lists and prints all Glue jobs in the current AWS account.
  def list_jobs
    @logger.info('Here are the Glue jobs in your account:')

    paginator = @client.get_jobs(max_results: 10)
    jobs = []

    paginator.each_page do |page|
      jobs.concat(page.jobs)
    end

    if jobs.empty?
      @logger.info("You don't have any Glue jobs.")
    else
      jobs.each do |job|
        @logger.info("- #{job.name}")
      end
    end
  end
end

if $PROGRAM_NAME == __FILE__
  glue_client = Aws::Glue::Client.new
  manager = GlueManager.new(glue_client)
  manager.list_jobs
end
```

- For API details, see ListJobs in *AWS SDK for Ruby API Reference*.

Rust

SDK for Rust

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        let mut list_jobs = glue.list_jobs().into_paginator().send();
        while let Some(list_jobs_output) = list_jobs.next().await {
            match list_jobs_output {
                Ok(list_jobs) => {
                    let names = list_jobs.job_names();
                    info!(?names, "Found these jobs")
                }
                Err(err) => return Err(GlueMvpError::from_glue_sdk(err)),
            }
        }
```

- For API details, see ListJobs in *AWS SDK for Rust API reference*.

AWS Glue API code examples

- Basic examples for AWS Glue using AWS SDKs
  - Hello AWS Glue
  - Learn the basics of AWS Glue with an AWS SDK
  - Actions for AWS Glue using AWS SDKs
    - Use CreateCrawler with an AWS SDK
    - Use CreateJob with an AWS SDK or CLI
    - Use DeleteCrawler with an AWS SDK
    - Use DeleteDatabase with an AWS SDK
    - Use DeleteJob with an AWS SDK or CLI
    - Use DeleteTable with an AWS SDK
    - Use GetCrawler with an AWS SDK

- Use GetDatabase with an AWS SDK
- Use GetDatabases with an AWS SDK or CLI
- Use GetJob with an AWS SDK or CLI
- Use GetJobRun with an AWS SDK or CLI
- Use GetJobRuns with an AWS SDK or CLI
- Use GetTables with an AWS SDK or CLI
- Use ListJobs with an AWS SDK
- Use StartCrawler with an AWS SDK or CLI
- Use StartJobRun with an AWS SDK or CLI

Basic examples for AWS Glue using AWS SDKs

The following code examples show how to use the basics of AWS Glue with AWS SDKs.

Examples

- Hello AWS Glue
- Learn the basics of AWS Glue with an AWS SDK
- Actions for AWS Glue using AWS SDKs
  - Use CreateCrawler with an AWS SDK
  - Use CreateJob with an AWS SDK or CLI
  - Use DeleteCrawler with an AWS SDK
  - Use DeleteDatabase with an AWS SDK
  - Use DeleteJob with an AWS SDK or CLI
  - Use DeleteTable with an AWS SDK
  - Use GetCrawler with an AWS SDK
  - Use GetDatabase with an AWS SDK
  - Use GetDatabases with an AWS SDK or CLI
  - Use GetJob with an AWS SDK or CLI
  - Use GetJobRun with an AWS SDK or CLI
  - Use GetJobRuns with an AWS SDK or CLI
  - Use GetTables with an AWS SDK or CLI

- Use ListJobs with an AWS SDK
- Use StartCrawler with an AWS SDK or CLI
- Use StartJobRun with an AWS SDK or CLI

Hello AWS Glue

The following code examples show how to get started using AWS Glue.

.NET

SDK for .NET

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
namespace GlueActions;

public class HelloGlue
{
    private static ILogger logger = null!;

    static async Task Main(string[] args)
    {
        // Set up dependency injection for AWS Glue.
        using var host = Host.CreateDefaultBuilder(args)
            .ConfigureLogging(logging =>
                logging.AddFilter("System", LogLevel.Debug)
                    .AddFilter<DebugLoggerProvider>("Microsoft",
 LogLevel.Information)
                    .AddFilter<ConsoleLoggerProvider>("Microsoft",
 LogLevel.Trace))
            .ConfigureServices((_, services) =>
                services.AddAWSService<IAmazonGlue>()
                .AddTransient<GlueWrapper>()
            )
            .Build();
        logger = LoggerFactory.Create(builder => { builder.AddConsole(); })
            .CreateLogger<HelloGlue>();
        var glueClient = host.Services.GetRequiredService<IAmazonGlue>();

        var request = new ListJobsRequest();

        var jobNames = new List<string>();

        do
        {
            var response = await glueClient.ListJobsAsync(request);
            jobNames.AddRange(response.JobNames);
            request.NextToken = response.NextToken;
        }
        while (request.NextToken is not null);

        Console.Clear();
        Console.WriteLine("Hello, Glue. Let's list your existing Glue Jobs:");
        if (jobNames.Count == 0)
        {
            Console.WriteLine("You don't have any AWS Glue jobs.");
        }
        else
        {
            jobNames.ForEach(Console.WriteLine);
        }
    }
}
```

- For API details, see ListJobs in *AWS SDK for .NET API Reference*.

C++

SDK for C++

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

Code for the CMakeLists.txt CMake file.

```
# Set the minimum required version of CMake for this project.
cmake_minimum_required(VERSION 3.13)

# Set the AWS service components used by this project.
set(SERVICE_COMPONENTS glue)

# Set this project's name.
project("hello_glue")

# Set the C++ standard to use to build this target.
# At least C++ 11 is required for the AWS SDK for C++.
set(CMAKE_CXX_STANDARD 11)

# Use the MSVC variable to determine if this is a Windows build.
set(WINDOWS_BUILD ${MSVC})

if (WINDOWS_BUILD) # Set the location where CMake can find the installed
 libraries for the AWS SDK.
    string(REPLACE ";" "/aws-cpp-sdk-all;" SYSTEM_MODULE_PATH
 "${CMAKE_SYSTEM_PREFIX_PATH}/aws-cpp-sdk-all")
    list(APPEND CMAKE_PREFIX_PATH ${SYSTEM_MODULE_PATH})
endif ()

# Find the AWS SDK for C++ package.
find_package(AWSSDK REQUIRED COMPONENTS ${SERVICE_COMPONENTS})

if (WINDOWS_BUILD AND AWSSDK_INSTALL_AS_SHARED_LIBS)
     # Copy relevant AWS SDK for C++ libraries into the current binary directory
 for running and debugging.

     # set(BIN_SUB_DIR "/Debug") # if you are building from the command line you
 may need to uncomment this
                                    # and set the proper subdirectory to the
 executables' location.

     AWSSDK_CPY_DYN_LIBS(SERVICE_COMPONENTS ""
 ${CMAKE_CURRENT_BINARY_DIR}${BIN_SUB_DIR})
endif ()

add_executable(${PROJECT_NAME}
        hello_glue.cpp)
```

```
target_link_libraries(${PROJECT_NAME}
        ${AWSSDK_LINK_LIBRARIES})
```

Code for the hello_glue.cpp source file.

```
#include <aws/core/Aws.h>
#include <aws/glue/GlueClient.h>
#include <aws/glue/model/ListJobsRequest.h>
#include <iostream>

/*
 *  A "Hello Glue" starter application which initializes an AWS Glue client and
 lists the
 *  AWS Glue job definitions.
 *
 *  main function
 *
 *  Usage: 'hello_glue'
 *
 */

int main(int argc, char **argv) {
    Aws::SDKOptions options;
    // Optionally change the log level for debugging.
//   options.loggingOptions.logLevel = Utils::Logging::LogLevel::Debug;
    Aws::InitAPI(options); // Should only be called once.
    int result = 0;
    {
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region (overrides config file).
        // clientConfig.region = "us-east-1";

        Aws::Glue::GlueClient glueClient(clientConfig);

        std::vector<Aws::String> jobs;

        Aws::String nextToken;  // Used for pagination.
        do {
            Aws::Glue::Model::ListJobsRequest listJobsRequest;
            if (!nextToken.empty()) {
                listJobsRequest.SetNextToken(nextToken);
            }
            Aws::Glue::Model::ListJobsOutcome listRunsOutcome =
 glueClient.ListJobs(
                    listJobsRequest);

            if (listRunsOutcome.IsSuccess()) {
                const std::vector<Aws::String> &jobNames =
 listRunsOutcome.GetResult().GetJobNames();
                jobs.insert(jobs.end(), jobNames.begin(), jobNames.end());

                nextToken = listRunsOutcome.GetResult().GetNextToken();
            } else {
                std::cerr << "Error listing jobs. "
                          << listRunsOutcome.GetError().GetMessage()
                          << std::endl;
                result = 1;
                break;
            }
        } while (!nextToken.empty());

        std::cout << "Your account has " << jobs.size() << " jobs."
                  << std::endl;
        for (size_t i = 0; i < jobs.size(); ++i) {
            std::cout << "   " << i + 1 << ". " << jobs[i] << std::endl;
        }
    }
    Aws::ShutdownAPI(options); // Should only be called once.
    return result;
}
```

- For API details, see ListJobs in *AWS SDK for C++ API Reference*.

Java

SDK for Java 2.x

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
package com.example.glue;

import software.amazon.awssdk.regions.Region;
import software.amazon.awssdk.services.glue.GlueClient;
import software.amazon.awssdk.services.glue.model.ListJobsRequest;
import software.amazon.awssdk.services.glue.model.ListJobsResponse;
import java.util.List;

public class HelloGlue {
    public static void main(String[] args) {
        GlueClient glueClient = GlueClient.builder()
                .region(Region.US_EAST_1)
                .build();

        listJobs(glueClient);
    }

    public static void listJobs(GlueClient glueClient) {
        ListJobsRequest request = ListJobsRequest.builder()
                .maxResults(10)
                .build();
        ListJobsResponse response = glueClient.listJobs(request);
        List<String> jobList = response.jobNames();
        jobList.forEach(job -> {
            System.out.println("Job Name: " + job);
        });
    }
}
```

- For API details, see ListJobs in *AWS SDK for Java 2.x API Reference*.

JavaScript

SDK for JavaScript (v3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
import { ListJobsCommand, GlueClient } from "@aws-sdk/client-glue";

const client = new GlueClient({});

export const main = async () => {
  const command = new ListJobsCommand({});

  const { JobNames } = await client.send(command);
  const formattedJobNames = JobNames.join("\n");
  console.log("Job names: ");
  console.log(formattedJobNames);
  return JobNames;
};
```

- For API details, see ListJobs in *AWS SDK for JavaScript API Reference*.

Python

SDK for Python (Boto3)

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```

import boto3
from botocore.exceptions import ClientError


def hello_glue():
    """
    Lists the job definitions in your AWS Glue account, using the AWS SDK for
 Python (Boto3).
    """
    try:
        # Create the Glue client
        glue = boto3.client("glue")
        # List the jobs, limiting the results to 10 per page
        paginator = glue.get_paginator("get_jobs")
        response_iterator = paginator.paginate(
            PaginationConfig={"MaxItems": 10, "PageSize": 10}
        )

        # Print the job names
        print("Here are the jobs in your account:")
        for page in response_iterator:
            for job in page["Jobs"]:
                print(f"\t{job['Name']}")

    except ClientError as e:
        print(f"Error: {e}")


if __name__ == "__main__":
    hello_glue()
```

- For API details, see ListJobs in *AWS SDK for Python (Boto3) API Reference*.

Ruby

SDK for Ruby

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
require 'aws-sdk-glue'
require 'logger'

# GlueManager is a class responsible for managing AWS Glue operations
# such as listing all Glue jobs in the current AWS account.
class GlueManager
  def initialize(client)
    @client = client
    @logger = Logger.new($stdout)
  end

  # Lists and prints all Glue jobs in the current AWS account.
  def list_jobs
    @logger.info('Here are the Glue jobs in your account:')

    paginator = @client.get_jobs(max_results: 10)
    jobs = []

    paginator.each_page do |page|
      jobs.concat(page.jobs)
    end

    if jobs.empty?
      @logger.info("You don't have any Glue jobs.")
    else
      jobs.each do |job|
        @logger.info("- #{job.name}")
      end
    end
  end
end

if $PROGRAM_NAME == __FILE__
  glue_client = Aws::Glue::Client.new
  manager = GlueManager.new(glue_client)
  manager.list_jobs
end
```

- For API details, see ListJobs in *AWS SDK for Ruby API Reference*.

Rust

SDK for Rust

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        let mut list_jobs = glue.list_jobs().into_paginator().send();
        while let Some(list_jobs_output) = list_jobs.next().await {
            match list_jobs_output {
                Ok(list_jobs) => {
                    let names = list_jobs.job_names();
                    info!(?names, "Found these jobs")
                }
                Err(err) => return Err(GlueMvpError::from_glue_sdk(err)),
            }
        }
```

- For API details, see ListJobs in *AWS SDK for Rust API reference*.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Learn the basics of AWS Glue with an AWS SDK

The following code examples show how to:

- Create a crawler that crawls a public Amazon S3 bucket and generates a database of CSV-formatted metadata.
- List information about databases and tables in your AWS Glue Data Catalog.
- Create a job to extract CSV data from the S3 bucket, transform the data, and load JSON-formatted output into another S3 bucket.
- List information about job runs, view transformed data, and clean up resources.

For more information, see Tutorial: Getting started with AWS Glue Studio.

.NET

SDK for .NET

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the [AWS Code Examples Repository](https://github.com/awsdocs/aws-doc-sdk-examples/tree/main/dotnetv3/Glue#code-examples).

Create a class that wraps AWS Glue functions that are used in the scenario.

```
using System.Net;

namespace GlueActions;

public class GlueWrapper
{
    private readonly IAmazonGlue _amazonGlue;

    /// <summary>
    /// Constructor for the AWS Glue actions wrapper.
    /// </summary>
    /// <param name="amazonGlue"></param>
    public GlueWrapper(IAmazonGlue amazonGlue)
    {
        _amazonGlue = amazonGlue;
    }

    /// <summary>
    /// Create an AWS Glue crawler.
    /// </summary>
    /// <param name="crawlerName">The name for the crawler.</param>
    /// <param name="crawlerDescription">A description of the crawler.</param>
    /// <param name="role">The AWS Identity and Access Management (IAM) role to
    /// be assumed by the crawler.</param>
    /// <param name="schedule">The schedule on which the crawler will be
 executed.</param>
    /// <param name="s3Path">The path to the Amazon Simple Storage Service
 (Amazon S3)
    /// bucket where the Python script has been stored.</param>
    /// <param name="dbName">The name to use for the database that will be
    /// created by the crawler.</param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> CreateCrawlerAsync(
        string crawlerName,
        string crawlerDescription,
        string role,
        string schedule,
        string s3Path,
        string dbName)
    {
        var s3Target = new S3Target
        {
            Path = s3Path,
        };

        var targetList = new List<S3Target>
        {
            s3Target,
        };

        var targets = new CrawlerTargets
        {
            S3Targets = targetList,
        };

        var crawlerRequest = new CreateCrawlerRequest
        {
            DatabaseName = dbName,
            Name = crawlerName,
            Description = crawlerDescription,
            Targets = targets,
            Role = role,
            Schedule = schedule,
        };

        var response = await _amazonGlue.CreateCrawlerAsync(crawlerRequest);
        return response.HttpStatusCode == System.Net.HttpStatusCode.OK;
    }


    /// <summary>
    /// Create an AWS Glue job.
    /// </summary>
    /// <param name="jobName">The name of the job.</param>
    /// <param name="roleName">The name of the IAM role to be assumed by
    /// the job.</param>
    /// <param name="description">A description of the job.</param>
    /// <param name="scriptUrl">The URL to the script.</param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> CreateJobAsync(string dbName, string tableName,
 string bucketUrl, string jobName, string roleName, string description, string
 scriptUrl)
    {
        var command = new JobCommand
        {
            PythonVersion = "3",
            Name = "glueetl",
            ScriptLocation = scriptUrl,
        };

        var arguments = new Dictionary<string, string>
        {
            { "--input_database", dbName },
            { "--input_table", tableName },
            { "--output_bucket_url", bucketUrl }
        };

        var request = new CreateJobRequest
        {
            Command = command,
            DefaultArguments = arguments,
            Description = description,
            GlueVersion = "3.0",
            Name = jobName,
            NumberOfWorkers = 10,
            Role = roleName,
            WorkerType = "G.1X"
        };

        var response = await _amazonGlue.CreateJobAsync(request);
        return response.HttpStatusCode == HttpStatusCode.OK;
    }


    /// <summary>
    /// Delete an AWS Glue crawler.
    /// </summary>
    /// <param name="crawlerName">The name of the crawler.</param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> DeleteCrawlerAsync(string crawlerName)
    {
        var response = await _amazonGlue.DeleteCrawlerAsync(new
 DeleteCrawlerRequest { Name = crawlerName });
        return response.HttpStatusCode == HttpStatusCode.OK;
    }


    /// <summary>
    /// Delete the AWS Glue database.
    /// </summary>
    /// <param name="dbName">The name of the database.</param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> DeleteDatabaseAsync(string dbName)
    {
        var response = await _amazonGlue.DeleteDatabaseAsync(new
 DeleteDatabaseRequest { Name = dbName });
        return response.HttpStatusCode == HttpStatusCode.OK;
    }


    /// <summary>
    /// Delete an AWS Glue job.
    /// </summary>
    /// <param name="jobName">The name of the job.</param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> DeleteJobAsync(string jobName)
    {
        var response = await _amazonGlue.DeleteJobAsync(new DeleteJobRequest
 { JobName = jobName });
        return response.HttpStatusCode == HttpStatusCode.OK;
    }


    /// <summary>
    /// Delete a table from an AWS Glue database.
    /// </summary>
    /// <param name="tableName">The table to delete.</param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> DeleteTableAsync(string dbName, string tableName)
    {
        var response = await _amazonGlue.DeleteTableAsync(new DeleteTableRequest
 { Name = tableName, DatabaseName = dbName });
```

```
        return response.HttpStatusCode == HttpStatusCode.OK;
    }


    /// <summary>
    /// Get information about an AWS Glue crawler.
    /// </summary>
    /// <param name="crawlerName">The name of the crawler.</param>
    /// <returns>A Crawler object describing the crawler.</returns>
    public async Task<Crawler?> GetCrawlerAsync(string crawlerName)
    {
        var crawlerRequest = new GetCrawlerRequest
        {
            Name = crawlerName,
        };

        var response = await _amazonGlue.GetCrawlerAsync(crawlerRequest);
        if (response.HttpStatusCode == System.Net.HttpStatusCode.OK)
        {
            var databaseName = response.Crawler.DatabaseName;
            Console.WriteLine($"{crawlerName} has the database {databaseName}");
            return response.Crawler;
        }

        Console.WriteLine($"No information regarding {crawlerName} could be
 found.");
        return null;
    }


    /// <summary>
    /// Get information about the state of an AWS Glue crawler.
    /// </summary>
    /// <param name="crawlerName">The name of the crawler.</param>
    /// <returns>A value describing the state of the crawler.</returns>
    public async Task<CrawlerState> GetCrawlerStateAsync(string crawlerName)
    {
        var response = await _amazonGlue.GetCrawlerAsync(
            new GetCrawlerRequest { Name = crawlerName });
        return response.Crawler.State;
    }


    /// <summary>
    /// Get information about an AWS Glue database.
    /// </summary>
    /// <param name="dbName">The name of the database.</param>
    /// <returns>A Database object containing information about the database.</
returns>
    public async Task<Database> GetDatabaseAsync(string dbName)
    {
        var databasesRequest = new GetDatabaseRequest
        {
            Name = dbName,
        };

        var response = await _amazonGlue.GetDatabaseAsync(databasesRequest);
        return response.Database;
    }


    /// <summary>
    /// Get information about a specific AWS Glue job run.
    /// </summary>
    /// <param name="jobName">The name of the job.</param>
    /// <param name="jobRunId">The Id of the job run.</param>
    /// <returns>A JobRun object with information about the job run.</returns>
    public async Task<JobRun> GetJobRunAsync(string jobName, string jobRunId)
    {
        var response = await _amazonGlue.GetJobRunAsync(new GetJobRunRequest
 { JobName = jobName, RunId = jobRunId });
        return response.JobRun;
    }


    /// <summary>
    /// Get information about all AWS Glue runs of a specific job.
    /// </summary>
    /// <param name="jobName">The name of the job.</param>
    /// <returns>A list of JobRun objects.</returns>
    public async Task<List<JobRun>> GetJobRunsAsync(string jobName)
    {
        var jobRuns = new List<JobRun>();

        var request = new GetJobRunsRequest
        {
            JobName = jobName,
        };
        // No need to loop to get all the log groups--the SDK does it for us
behind the scenes
        var paginatorForJobRuns =
            _amazonGlue.Paginators.GetJobRuns(request);

        await foreach (var response in paginatorForJobRuns.Responses)
        {
            response.JobRuns.ForEach(jobRun =>
            {
                jobRuns.Add(jobRun);
            });
        }

        return jobRuns;
    }


    /// <summary>
    /// Get a list of tables for an AWS Glue database.
    /// </summary>
    /// <param name="dbName">The name of the database.</param>
    /// <returns>A list of Table objects.</returns>
    public async Task<List<Table>> GetTablesAsync(string dbName)
    {
        var request = new GetTablesRequest { DatabaseName = dbName };
        var tables = new List<Table>();

        // Get a paginator for listing the tables.
        var tablePaginator = _amazonGlue.Paginators.GetTables(request);

        await foreach (var response in tablePaginator.Responses)
        {
            tables.AddRange(response.TableList);
        }

        return tables;
    }


    /// <summary>
    /// List AWS Glue jobs using a paginator.
    /// </summary>
    /// <returns>A list of AWS Glue job names.</returns>
    public async Task<List<string>> ListJobsAsync()
    {
        var jobNames = new List<string>();

        var listJobsPaginator = _amazonGlue.Paginators.ListJobs(new
 ListJobsRequest { MaxResults = 10 });
        await foreach (var response in listJobsPaginator.Responses)
        {
            jobNames.AddRange(response.JobNames);
        }

        return jobNames;
    }


    /// <summary>
    /// Start an AWS Glue crawler.
    /// </summary>
    /// <param name="crawlerName">The name of the crawler.</param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> StartCrawlerAsync(string crawlerName)
    {
        var crawlerRequest = new StartCrawlerRequest
        {
            Name = crawlerName,
        };

        var response = await _amazonGlue.StartCrawlerAsync(crawlerRequest);

        return response.HttpStatusCode == System.Net.HttpStatusCode.OK;
    }


    /// <summary>
    /// Start an AWS Glue job run.
    /// </summary>
    /// <param name="jobName">The name of the job.</param>
    /// <returns>A string representing the job run Id.</returns>
    public async Task<string> StartJobRunAsync(
        string jobName,
        string inputDatabase,
        string inputTable,
        string bucketName)
    {
        var request = new StartJobRunRequest
        {
            JobName = jobName,
            Arguments = new Dictionary<string, string>
            {
                {"--input_database", inputDatabase},
                {"--input_table", inputTable},
                {"--output_bucket_url", $"s3://{bucketName}/"}
            }
        };

        var response = await _amazonGlue.StartJobRunAsync(request);
        return response.JobRunId;
    }

}
```

Create a class that runs the scenario.

```
global using Amazon.Glue;
global using GlueActions;
global using Microsoft.Extensions.Configuration;
global using Microsoft.Extensions.DependencyInjection;
global using Microsoft.Extensions.Hosting;
global using Microsoft.Extensions.Logging;
global using Microsoft.Extensions.Logging.Console;
global using Microsoft.Extensions.Logging.Debug;



using Amazon.Glue.Model;
using Amazon.S3;
using Amazon.S3.Model;

namespace GlueBasics;

public class GlueBasics
{
    private static ILogger logger = null!;
    private static IConfiguration _configuration = null!;
    static async Task Main(string[] args)
    {
        // Set up dependency injection for AWS Glue.
        using var host = Host.CreateDefaultBuilder(args)
            .ConfigureLogging(logging =>
                logging.AddFilter("System", LogLevel.Debug)
                    .AddFilter<DebugLoggerProvider>("Microsoft",
 LogLevel.Information)
                    .AddFilter<ConsoleLoggerProvider>("Microsoft",
 LogLevel.Trace))
            .ConfigureServices((_, services) =>
            services.AddAWSService<IAmazonGlue>()
            .AddTransient<GlueWrapper>()
            .AddTransient<UiWrapper>()
            )
            .Build();

        logger = LoggerFactory.Create(builder => { builder.AddConsole(); })
        .CreateLogger<GlueBasics>();

        _configuration = new ConfigurationBuilder()
            .SetBasePath(Directory.GetCurrentDirectory())
            .AddJsonFile("settings.json") // Load settings from .json file.
            .AddJsonFile("settings.local.json",
                true) // Optionally load local settings.
            .Build();

        // These values are stored in settings.json
        // Once you have run the CDK script to deploy the resources,
        // edit the file to set "BucketName", "RoleName", and "ScriptURL"
        // to the appropriate values. Also set "CrawlerName" to the name
        // you want to give the crawler when it is created.
        string bucketName = _configuration["BucketName"]!;
        string bucketUrl = _configuration["BucketUrl"]!;
        string crawlerName = _configuration["CrawlerName"]!;
        string roleName = _configuration["RoleName"]!;
        string sourceData = _configuration["SourceData"]!;
        string dbName = _configuration["DbName"]!;
        string cron = _configuration["Cron"]!;
        string scriptUrl = _configuration["ScriptURL"]!;
        string jobName = _configuration["JobName"]!;

        var wrapper = host.Services.GetRequiredService<GlueWrapper>();
        var uiWrapper = host.Services.GetRequiredService<UiWrapper>();
        uiWrapper.DisplayOverview();
        uiWrapper.PressEnter();

        // Create the crawler and wait for it to be ready.
        uiWrapper.DisplayTitle("Create AWS Glue crawler");
        Console.WriteLine("Let's begin by creating the AWS Glue crawler.");

        var crawlerDescription = "Crawler created for the AWS Glue Basics 
scenario.";
        var crawlerCreated = await wrapper.CreateCrawlerAsync(crawlerName, 
crawlerDescription, roleName, cron, sourceData, dbName);
        if (crawlerCreated)
        {
            Console.WriteLine($"The crawler: {crawlerName} has been created. Now 
let's wait until it's ready.");
            CrawlerState crawlerState;
            do
            {
                crawlerState = await wrapper.GetCrawlerStateAsync(crawlerName);
            }
            while (crawlerState != "READY");
            Console.WriteLine($"The crawler {crawlerName} is now ready for 
use.");
        }
        else
        {
            Console.WriteLine($"Couldn't create crawler {crawlerName}.");
            return; // Exit the application.
        }

        uiWrapper.DisplayTitle("Start AWS Glue crawler");
        Console.WriteLine("Now let's wait until the crawler has successfully 
started.");
        var crawlerStarted = await wrapper.StartCrawlerAsync(crawlerName);
        if (crawlerStarted)
        {
            CrawlerState crawlerState;
            do
            {
                crawlerState = await wrapper.GetCrawlerStateAsync(crawlerName);
            }
            while (crawlerState != "READY");
            Console.WriteLine($"The crawler {crawlerName} is now ready for 
use.");
        }
        else
        {
            Console.WriteLine($"Couldn't start the crawler {crawlerName}.");
            return; // Exit the application.
        }

        uiWrapper.PressEnter();

        Console.WriteLine($"\nLet's take a look at the database: {dbName}");
        var database = await wrapper.GetDatabaseAsync(dbName);

        if (database != null)
        {
            uiWrapper.DisplayTitle($"{database.Name} Details");
            Console.WriteLine($"{database.Name} created on 
{database.CreateTime}");
            Console.WriteLine(database.Description);
        }

        uiWrapper.PressEnter();

        var tables = await wrapper.GetTablesAsync(dbName);
        if (tables.Count > 0)
        {
            tables.ForEach(table =>
            {
                Console.WriteLine($"{table.Name}\tCreated: 
{table.CreateTime}\tUpdated: {table.UpdateTime}");
            });
        }

        uiWrapper.PressEnter();

        uiWrapper.DisplayTitle("Create AWS Glue job");
        Console.WriteLine("Creating a new AWS Glue job.");
        var description = "An AWS Glue job created using the AWS SDK for .NET";
        await wrapper.CreateJobAsync(dbName, tables[0].Name, bucketUrl, jobName, 
roleName, description, scriptUrl);

        uiWrapper.PressEnter();
        uiWrapper.DisplayTitle("Starting AWS Glue job");
        Console.WriteLine("Starting the new AWS Glue job...");
        var jobRunId = await wrapper.StartJobRunAsync(jobName, dbName,
 tables[0].Name, bucketName);
        var jobRunComplete = false;
        var jobRun = new JobRun();
        do
        {
            jobRun = await wrapper.GetJobRunAsync(jobName, jobRunId);
            if (jobRun.JobRunState == "SUCCEEDED" || jobRun.JobRunState ==
 "STOPPED" ||
                jobRun.JobRunState == "FAILED" || jobRun.JobRunState ==
 "TIMEOUT")
            {
                jobRunComplete = true;
            }
        } while (!jobRunComplete);

        uiWrapper.DisplayTitle($"Data in {bucketName}");

        // Get the list of data stored in the S3 bucket.
        var s3Client = new AmazonS3Client();

        var response = await s3Client.ListObjectsAsync(new ListObjectsRequest
 { BucketName = bucketName });
        response.S3Objects.ForEach(s3Object =>
        {
            Console.WriteLine(s3Object.Key);
        });

        uiWrapper.DisplayTitle("AWS Glue jobs");
        var jobNames = await wrapper.ListJobsAsync();
        jobNames.ForEach(jobName =>
        {
            Console.WriteLine(jobName);
        });

        uiWrapper.PressEnter();

        uiWrapper.DisplayTitle("Get AWS Glue job run information");
        Console.WriteLine("Getting information about the AWS Glue job.");
        var jobRuns = await wrapper.GetJobRunsAsync(jobName);

        jobRuns.ForEach(jobRun =>
        {

  Console.WriteLine($"{jobRun.JobName}\t{jobRun.JobRunState}\t{jobRun.CompletedOn}");
        });

        uiWrapper.PressEnter();

        uiWrapper.DisplayTitle("Deleting resources");
        Console.WriteLine("Deleting the AWS Glue job used by the example.");
        await wrapper.DeleteJobAsync(jobName);

        Console.WriteLine("Deleting the tables from the database.");
        tables.ForEach(async table =>
        {
            await wrapper.DeleteTableAsync(dbName, table.Name);
        });

        Console.WriteLine("Deleting the database.");
        await wrapper.DeleteDatabaseAsync(dbName);

        Console.WriteLine("Deleting the AWS Glue crawler.");
        await wrapper.DeleteCrawlerAsync(crawlerName);

        Console.WriteLine("The AWS Glue scenario has completed.");
        uiWrapper.PressEnter();
    }
}


namespace GlueBasics;

public class UiWrapper
{
    public readonly string SepBar = new string('-', Console.WindowWidth);

    /// <summary>
    /// Show information about the scenario.
    /// </summary>
    public void DisplayOverview()
    {
        Console.Clear();
        DisplayTitle("Amazon Glue: get started with crawlers and jobs");

        Console.WriteLine("This example application does the following:");
        Console.WriteLine("\t 1. Create a crawler, pass it the IAM role and the
 URL to the public S3 bucket that contains the source data");
        Console.WriteLine("\t 2. Start the crawler.");
        Console.WriteLine("\t 3. Get the database created by the crawler and the
 tables in the database.");
        Console.WriteLine("\t 4. Create a job.");
        Console.WriteLine("\t 5. Start a job run.");
        Console.WriteLine("\t 6. Wait for the job run to complete.");
        Console.WriteLine("\t 7. Show the data stored in the bucket.");
        Console.WriteLine("\t 8. List jobs for the account.");
        Console.WriteLine("\t 9. Get job run details for the job that was run.");
        Console.WriteLine("\t10. Delete the demo job.");
        Console.WriteLine("\t11. Delete the database and tables created for the
 demo.");
        Console.WriteLine("\t12. Delete the crawler.");
    }

    /// <summary>
    /// Display a message and wait until the user presses enter.
    /// </summary>
    public void PressEnter()
    {
        Console.Write("\nPlease press <Enter> to continue. ");
        _ = Console.ReadLine();
    }

    /// <summary>
    /// Pad a string with spaces to center it on the console display.
    /// </summary>
    /// <param name="strToCenter">The string to center on the screen.</param>
    /// <returns>The string padded to make it center on the screen.</returns>
    public string CenterString(string strToCenter)
    {
        var padAmount = (Console.WindowWidth - strToCenter.Length) / 2;
        var leftPad = new string(' ', padAmount);
        return $"{leftPad}{strToCenter}";
    }

    /// <summary>
    /// Display a line of hyphens, the centered text of the title and another
    /// line of hyphens.
    /// </summary>
    /// <param name="strTitle">The string to be displayed.</param>
    public void DisplayTitle(string strTitle)
    {
        Console.WriteLine(SepBar);
        Console.WriteLine(CenterString(strTitle));
        Console.WriteLine(SepBar);
    }
}
```

- For API details, see the following topics in *AWS SDK for .NET API Reference*.
  - CreateCrawler
  - CreateJob
  - DeleteCrawler
  - DeleteDatabase
  - DeleteJob
  - DeleteTable
  - GetCrawler
  - GetDatabase
  - GetDatabases
  - GetJob
  - GetJobRun
  - GetJobRuns
  - GetTables
  - ListJobs
  - StartCrawler
  - StartJobRun

C++

SDK for C++

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
//! Scenario which demonstrates using AWS Glue to add a crawler and run a job.
/*!
 \\sa runGettingStartedWithGlueScenario()
 \param bucketName: An S3 bucket created in the setup.
 \param roleName: An AWS Identity and Access Management (IAM) role created in the
 setup.
 \param clientConfig: AWS client configuration.
 \return bool: Successful completion.
 */

bool AwsDoc::Glue::runGettingStartedWithGlueScenario(const Aws::String
 &bucketName,
                                                     const Aws::String &roleName,
                                                     const
 Aws::Client::ClientConfiguration &clientConfig) {
    Aws::Glue::GlueClient client(clientConfig);

    Aws::String roleArn;
    if (!getRoleArn(roleName, roleArn, clientConfig)) {
        std::cerr << "Error getting role ARN for role." << std::endl;
        return false;
    }

    // 1. Upload the job script to the S3 bucket.
    {
        std::cout << "Uploading the job script '"
                  << AwsDoc::Glue::PYTHON_SCRIPT
                  << "'." << std::endl;

        if (!AwsDoc::Glue::uploadFile(bucketName,
                                      AwsDoc::Glue::PYTHON_SCRIPT_PATH,
                                      AwsDoc::Glue::PYTHON_SCRIPT,
                                       clientConfig)) {
                std::cerr << "Error uploading the job file." << std::endl;
                return false;
            }
        }

        // 2. Create a crawler.
        {
            Aws::Glue::Model::S3Target s3Target;
            s3Target.SetPath("s3://crawler-public-us-east-1/flight/2016/csv");
            Aws::Glue::Model::CrawlerTargets crawlerTargets;
            crawlerTargets.AddS3Targets(s3Target);

            Aws::Glue::Model::CreateCrawlerRequest request;
            request.SetTargets(crawlerTargets);
            request.SetName(CRAWLER_NAME);
            request.SetDatabaseName(CRAWLER_DATABASE_NAME);
            request.SetTablePrefix(CRAWLER_DATABASE_PREFIX);
            request.SetRole(roleArn);

            Aws::Glue::Model::CreateCrawlerOutcome outcome =
 client.CreateCrawler(request);

            if (outcome.IsSuccess()) {
                std::cout << "Successfully created the crawler." << std::endl;
            }
            else {
                std::cerr << "Error creating a crawler. " <<
 outcome.GetError().GetMessage()
                          << std::endl;
                deleteAssets("", CRAWLER_DATABASE_NAME, "", bucketName,
 clientConfig);
                return false;
            }
        }

        // 3. Get a crawler.
        {
            Aws::Glue::Model::GetCrawlerRequest request;
            request.SetName(CRAWLER_NAME);

            Aws::Glue::Model::GetCrawlerOutcome outcome = client.GetCrawler(request);

            if (outcome.IsSuccess()) {
```

```
           Aws::Glue::Model::CrawlerState crawlerState =
outcome.GetResult().GetCrawler().GetState();
           std::cout << "Retrieved crawler with state " <<

Aws::Glue::Model::CrawlerStateMapper::GetNameForCrawlerState(
                             crawlerState)
                     << "." << std::endl;
       }
       else {
           std::cerr << "Error retrieving a crawler.  "
                     << outcome.GetError().GetMessage() << std::endl;
           deleteAssets(CRAWLER_NAME, CRAWLER_DATABASE_NAME, "", bucketName,
                        clientConfig);
           return false;
       }
   }

   // 4. Start a crawler.
   {
       Aws::Glue::Model::StartCrawlerRequest request;
       request.SetName(CRAWLER_NAME);

       Aws::Glue::Model::StartCrawlerOutcome outcome =
client.StartCrawler(request);


       if (outcome.IsSuccess() || (Aws::Glue::GlueErrors::CRAWLER_RUNNING ==
                                   outcome.GetError().GetErrorType())) {
           if (!outcome.IsSuccess()) {
               std::cout << "Crawler was already started." << std::endl;
           }
           else {
               std::cout << "Successfully started crawler." << std::endl;
           }

           std::cout << "This may take a while to run." << std::endl;

           Aws::Glue::Model::CrawlerState crawlerState =
Aws::Glue::Model::CrawlerState::NOT_SET;
           int iterations = 0;
           while (Aws::Glue::Model::CrawlerState::READY != crawlerState) {
               std::this_thread::sleep_for(std::chrono::seconds(1));
               ++iterations;
               if ((iterations % 10) == 0) { // Log status every 10 seconds.
```

```
                    std::cout << "Crawler status " <<

Aws::Glue::Model::CrawlerStateMapper::GetNameForCrawlerState(
                                     crawlerState)
                             << ". After " << iterations
                             << " seconds elapsed."
                             << std::endl;
                }
                Aws::Glue::Model::GetCrawlerRequest getCrawlerRequest;
                getCrawlerRequest.SetName(CRAWLER_NAME);

                Aws::Glue::Model::GetCrawlerOutcome getCrawlerOutcome =
client.GetCrawler(
                        getCrawlerRequest);

                if (getCrawlerOutcome.IsSuccess()) {
                    crawlerState =
getCrawlerOutcome.GetResult().GetCrawler().GetState();
                }
                else {
                    std::cerr << "Error getting crawler.  "
                              << getCrawlerOutcome.GetError().GetMessage() <<
std::endl;
                    break;
                }
            }

            if (Aws::Glue::Model::CrawlerState::READY == crawlerState) {
                std::cout << "Crawler finished running after " << iterations
                          << " seconds."
                          << std::endl;
            }
        }
        else {
            std::cerr << "Error starting a crawler.  "
                      << outcome.GetError().GetMessage()
                      << std::endl;

            deleteAssets(CRAWLER_NAME, CRAWLER_DATABASE_NAME, "", bucketName,
                         clientConfig);
            return false;
        }
    }
    // 5. Get a database.
    {
        Aws::Glue::Model::GetDatabaseRequest request;
        request.SetName(CRAWLER_DATABASE_NAME);

        Aws::Glue::Model::GetDatabaseOutcome outcome =
 client.GetDatabase(request);

        if (outcome.IsSuccess()) {
            const Aws::Glue::Model::Database &database =
 outcome.GetResult().GetDatabase();

            std::cout << "Successfully retrieve the database\n" <<
                      database.Jsonize().View().WriteReadable() << "'." <<
 std::endl;
        }
        else {
            std::cerr << "Error getting the database.  "
                      << outcome.GetError().GetMessage() << std::endl;
            deleteAssets(CRAWLER_NAME, CRAWLER_DATABASE_NAME, "", bucketName,
                         clientConfig);
            return false;
        }
    }

    // 6. Get tables.
    Aws::String tableName;
    {
        Aws::Glue::Model::GetTablesRequest request;
        request.SetDatabaseName(CRAWLER_DATABASE_NAME);
        std::vector<Aws::Glue::Model::Table> all_tables;
        Aws::String nextToken; // Used for pagination.
        do {
            Aws::Glue::Model::GetTablesOutcome outcome =
 client.GetTables(request);

            if (outcome.IsSuccess()) {
                const std::vector<Aws::Glue::Model::Table> &tables =
 outcome.GetResult().GetTableList();
                all_tables.insert(all_tables.end(), tables.begin(),
 tables.end());
                nextToken = outcome.GetResult().GetNextToken();
            }
            else {
                std::cerr << "Error getting the tables. "
                          << outcome.GetError().GetMessage()
                          << std::endl;
                deleteAssets(CRAWLER_NAME, CRAWLER_DATABASE_NAME, "", bucketName,
                             clientConfig);
                return false;
            }
        } while (!nextToken.empty());

        std::cout << "The database contains " << all_tables.size()
                  << (all_tables.size() == 1 ?
                      " table." : "tables.") << std::endl;
        std::cout << "Here is a list of the tables in the database.";
        for (size_t index = 0; index < all_tables.size(); ++index) {
            std::cout << "    " << index + 1 << ":  " <<
 all_tables[index].GetName()
                      << std::endl;
        }

        if (!all_tables.empty()) {
            int tableIndex = askQuestionForIntRange(
                    "Enter an index to display the database detail ",
                    1, static_cast<int>(all_tables.size()));
            std::cout << all_tables[tableIndex -
 1].Jsonize().View().WriteReadable()
                      << std::endl;

            tableName = all_tables[tableIndex - 1].GetName();
        }
    }

    // 7. Create a job.
    {
        Aws::Glue::Model::CreateJobRequest request;
        request.SetName(JOB_NAME);
        request.SetRole(roleArn);
        request.SetGlueVersion(GLUE_VERSION);

        Aws::Glue::Model::JobCommand command;
        command.SetName(JOB_COMMAND_NAME);
        command.SetPythonVersion(JOB_PYTHON_VERSION);
        command.SetScriptLocation(
                Aws::String("s3://") + bucketName + "/" + PYTHON_SCRIPT);
        request.SetCommand(command);
        Aws::Glue::Model::CreateJobOutcome outcome = client.CreateJob(request);

        if (outcome.IsSuccess()) {
            std::cout << "Successfully created the job." << std::endl;
        }
        else {
            std::cerr << "Error creating the job. " <<
outcome.GetError().GetMessage()
                      << std::endl;
            deleteAssets(CRAWLER_NAME, CRAWLER_DATABASE_NAME, "", bucketName,
                         clientConfig);
            return false;
        }
    }

    // 8. Start a job run.
    {
        Aws::Glue::Model::StartJobRunRequest request;
        request.SetJobName(JOB_NAME);

        Aws::Map<Aws::String, Aws::String> arguments;
        arguments["--input_database"] = CRAWLER_DATABASE_NAME;
        arguments["--input_table"] = tableName;
        arguments["--output_bucket_url"] = Aws::String("s3://") + bucketName +
"/";
        request.SetArguments(arguments);

        Aws::Glue::Model::StartJobRunOutcome outcome =
client.StartJobRun(request);

        if (outcome.IsSuccess()) {
            std::cout << "Successfully started the job." << std::endl;

            Aws::String jobRunId = outcome.GetResult().GetJobRunId();

            int iterator = 0;
            bool done = false;
            while (!done) {
                ++iterator;
                std::this_thread::sleep_for(std::chrono::seconds(1));
                Aws::Glue::Model::GetJobRunRequest jobRunRequest;
                jobRunRequest.SetJobName(JOB_NAME);
                jobRunRequest.SetRunId(jobRunId);
                Aws::Glue::Model::GetJobRunOutcome jobRunOutcome =
client.GetJobRun(
                        jobRunRequest);

                if (jobRunOutcome.IsSuccess()) {
                    const Aws::Glue::Model::JobRun &jobRun =
jobRunOutcome.GetResult().GetJobRun();
                    Aws::Glue::Model::JobRunState jobRunState =
jobRun.GetJobRunState();

                    if ((jobRunState == Aws::Glue::Model::JobRunState::STOPPED)
||
                        (jobRunState == Aws::Glue::Model::JobRunState::FAILED) ||
                        (jobRunState == Aws::Glue::Model::JobRunState::TIMEOUT))
{
                        std::cerr << "Error running job. "
                                  << jobRun.GetErrorMessage()
                                  << std::endl;
                        deleteAssets(CRAWLER_NAME, CRAWLER_DATABASE_NAME,
JOB_NAME,
                                     bucketName,
                                     clientConfig);
                        return false;
                    }
                    else if (jobRunState ==
                             Aws::Glue::Model::JobRunState::SUCCEEDED) {
                        std::cout << "Job run succeeded after  " << iterator <<
                                  " seconds elapsed." << std::endl;
                        done = true;
                    }
                    else if ((iterator % 10) == 0) { // Log status every 10
seconds.
                        std::cout << "Job run status " <<

Aws::Glue::Model::JobRunStateMapper::GetNameForJobRunState(
                                          jobRunState) <<
                                  ". " << iterator <<
                                  " seconds elapsed." << std::endl;
                    }
                }
                else {
                    std::cerr << "Error retrieving job run state. "
                              << jobRunOutcome.GetError().GetMessage()
                             << std::endl;
                   deleteAssets(CRAWLER_NAME, CRAWLER_DATABASE_NAME, JOB_NAME,
                                bucketName, clientConfig);
                   return false;
               }
           }
       }
       else {
           std::cerr << "Error starting a job. " <<
outcome.GetError().GetMessage()
                     << std::endl;
           deleteAssets(CRAWLER_NAME, CRAWLER_DATABASE_NAME, JOB_NAME,
bucketName,
                        clientConfig);
           return false;
       }
   }

   // 9. List the output data stored in the S3 bucket.
   {
       Aws::S3::S3Client s3Client;
       Aws::S3::Model::ListObjectsV2Request request;
       request.SetBucket(bucketName);
       request.SetPrefix(OUTPUT_FILE_PREFIX);

       Aws::String continuationToken; // Used for pagination.
       std::vector<Aws::S3::Model::Object> allObjects;
       do {
           if (!continuationToken.empty()) {
               request.SetContinuationToken(continuationToken);
           }
           Aws::S3::Model::ListObjectsV2Outcome outcome =
s3Client.ListObjectsV2(
                   request);

           if (outcome.IsSuccess()) {
               const std::vector<Aws::S3::Model::Object> &objects =
                       outcome.GetResult().GetContents();
               allObjects.insert(allObjects.end(), objects.begin(),
objects.end());
               continuationToken =
outcome.GetResult().GetNextContinuationToken();
           }
           else {
                std::cerr << "Error listing objects. "
                          << outcome.GetError().GetMessage()
                          << std::endl;
                break;
            }
        } while (!continuationToken.empty());

        std::cout << "Data from your job is in " << allObjects.size() <<
                  " files in the S3 bucket, " << bucketName << "." << std::endl;

        for (size_t i = 0; i < allObjects.size(); ++i) {
            std::cout << "    " << i + 1 << ". " << allObjects[i].GetKey()
                      << std::endl;
        }

        int objectIndex = askQuestionForIntRange(
                std::string(
                        "Enter the number of a block to download it and see the
 first ") +
                std::to_string(LINES_OF_RUN_FILE_TO_DISPLAY) +
                " lines of JSON output in the block: ", 1,
                static_cast<int>(allObjects.size()));

        Aws::String objectKey = allObjects[objectIndex - 1].GetKey();

        std::stringstream stringStream;
        if (getObjectFromBucket(bucketName, objectKey, stringStream,
                                clientConfig)) {
            for (int i = 0; i < LINES_OF_RUN_FILE_TO_DISPLAY && stringStream; +
+i) {
                std::string line;
                std::getline(stringStream, line);
                std::cout << "    " << line << std::endl;
            }
        }
        else {
            deleteAssets(CRAWLER_NAME, CRAWLER_DATABASE_NAME, JOB_NAME,
 bucketName,
                         clientConfig);
            return false;
        }
    }

    // 10. List all the jobs.
    Aws::String jobName;
    {
        Aws::Glue::Model::ListJobsRequest listJobsRequest;

        Aws::String nextToken;
        std::vector<Aws::String> allJobNames;

        do {
            if (!nextToken.empty()) {
                listJobsRequest.SetNextToken(nextToken);
            }
            Aws::Glue::Model::ListJobsOutcome listRunsOutcome = client.ListJobs(
                    listJobsRequest);

            if (listRunsOutcome.IsSuccess()) {
                const std::vector<Aws::String> &jobNames =
 listRunsOutcome.GetResult().GetJobNames();
                allJobNames.insert(allJobNames.end(), jobNames.begin(),
 jobNames.end());
                nextToken = listRunsOutcome.GetResult().GetNextToken();
            }
            else {
                std::cerr << "Error listing jobs. "
                          << listRunsOutcome.GetError().GetMessage()
                          << std::endl;
            }
        } while (!nextToken.empty());
        std::cout << "Your account has " << allJobNames.size() << " jobs."
                  << std::endl;
        for (size_t i = 0; i < allJobNames.size(); ++i) {
            std::cout << "   " << i + 1 << ". " << allJobNames[i] << std::endl;
        }
        int jobIndex = askQuestionForIntRange(
                Aws::String("Enter a number between 1 and ") +
                std::to_string(allJobNames.size()) +
                " to see the list of runs for a job: ",
                1, static_cast<int>(allJobNames.size()));

        jobName = allJobNames[jobIndex - 1];
    }

    // 11. Get the job runs for a job.
    Aws::String jobRunID;
    if (!jobName.empty()) {
        Aws::Glue::Model::GetJobRunsRequest getJobRunsRequest;
        getJobRunsRequest.SetJobName(jobName);

        Aws::String nextToken; // Used for pagination.
        std::vector<Aws::Glue::Model::JobRun> allJobRuns;
        do {
            if (!nextToken.empty()) {
                getJobRunsRequest.SetNextToken(nextToken);
            }
            Aws::Glue::Model::GetJobRunsOutcome jobRunsOutcome =
 client.GetJobRuns(
                    getJobRunsRequest);

            if (jobRunsOutcome.IsSuccess()) {
                const std::vector<Aws::Glue::Model::JobRun> &jobRuns =
 jobRunsOutcome.GetResult().GetJobRuns();
                allJobRuns.insert(allJobRuns.end(), jobRuns.begin(),
 jobRuns.end());

                nextToken = jobRunsOutcome.GetResult().GetNextToken();
            }
            else {
                std::cerr << "Error getting job runs. "
                          << jobRunsOutcome.GetError().GetMessage()
                          << std::endl;
                break;
            }
        } while (!nextToken.empty());

        std::cout << "There are " << allJobRuns.size() << " runs in the job '"
                  <<
                  jobName << "'." << std::endl;

        for (size_t i = 0; i < allJobRuns.size(); ++i) {
            std::cout << "   " << i + 1 << ". " << allJobRuns[i].GetJobName()
                      << std::endl;
        }

        int runIndex = askQuestionForIntRange(
                Aws::String("Enter a number between 1 and ") +
                std::to_string(allJobRuns.size()) +
                " to see details for a run: ",
                1, static_cast<int>(allJobRuns.size()));
        jobRunID = allJobRuns[runIndex - 1].GetId();
```

```
    }

    // 12. Get a single job run.
    if (!jobRunID.empty()) {
        Aws::Glue::Model::GetJobRunRequest jobRunRequest;
        jobRunRequest.SetJobName(jobName);
        jobRunRequest.SetRunId(jobRunID);

        Aws::Glue::Model::GetJobRunOutcome jobRunOutcome = client.GetJobRun(
                jobRunRequest);

        if (jobRunOutcome.IsSuccess()) {
            std::cout << "Displaying the job run JSON description." << std::endl;
            std::cout
                    <<
 jobRunOutcome.GetResult().GetJobRun().Jsonize().View().WriteReadable()
                    << std::endl;
        }
        else {
            std::cerr << "Error get a job run. "
                      << jobRunOutcome.GetError().GetMessage()
                      << std::endl;
        }
    }

    return deleteAssets(CRAWLER_NAME, CRAWLER_DATABASE_NAME, JOB_NAME,
 bucketName,
                        clientConfig);
}

//! Cleanup routine to delete created assets.
/*!
 \\sa deleteAssets()
 \param crawler: Name of an AWS Glue crawler.
 \param database: The name of an AWS Glue database.
 \param job: The name of an AWS Glue job.
 \param bucketName: The name of an S3 bucket.
 \param clientConfig: AWS client configuration.
 \return bool: Successful completion.
 */
bool AwsDoc::Glue::deleteAssets(const Aws::String &crawler, const Aws::String
 &database,
                                const Aws::String &job, const Aws::String
 &bucketName,
                                const Aws::Client::ClientConfiguration
&clientConfig) {
   const Aws::Glue::GlueClient client(clientConfig);
   bool result = true;

   // 13. Delete a job.
   if (!job.empty()) {
       Aws::Glue::Model::DeleteJobRequest request;
       request.SetJobName(job);

       Aws::Glue::Model::DeleteJobOutcome outcome = client.DeleteJob(request);


       if (outcome.IsSuccess()) {
           std::cout << "Successfully deleted the job." << std::endl;
       }
       else {
           std::cerr << "Error deleting the job. " <<
outcome.GetError().GetMessage()
                     << std::endl;
           result = false;
       }
   }

   // 14. Delete a database.
   if (!database.empty()) {
       Aws::Glue::Model::DeleteDatabaseRequest request;
       request.SetName(database);

       Aws::Glue::Model::DeleteDatabaseOutcome outcome = client.DeleteDatabase(
               request);

       if (outcome.IsSuccess()) {
           std::cout << "Successfully deleted the database." << std::endl;
       }
       else {
           std::cerr << "Error deleting database. " <<
outcome.GetError().GetMessage()
                     << std::endl;
           result = false;
       }
   }

   // 15. Delete a crawler.
```

```
    if (!crawler.empty()) {
        Aws::Glue::Model::DeleteCrawlerRequest request;
        request.SetName(crawler);

        Aws::Glue::Model::DeleteCrawlerOutcome outcome =
 client.DeleteCrawler(request);

        if (outcome.IsSuccess()) {
            std::cout << "Successfully deleted the crawler." << std::endl;
        }
        else {
            std::cerr << "Error deleting the crawler. "
                      << outcome.GetError().GetMessage() << std::endl;
            result = false;
        }
    }

    // 16. Delete the job script and run data from the S3 bucket.
    result &= AwsDoc::Glue::deleteAllObjectsInS3Bucket(bucketName,
                                                        clientConfig);
    return result;
}

//! Routine which uploads a file to an S3 bucket.
/*!
 \\sa uploadFile()
 \param bucketName: An S3 bucket created in the setup.
 \param filePath: The path of the file to upload.
 \param fileName The name for the uploaded file.
 \param clientConfig: AWS client configuration.
 \return bool: Successful completion.
 */
bool
AwsDoc::Glue::uploadFile(const Aws::String &bucketName,
                         const Aws::String &filePath,
                         const Aws::String &fileName,
                         const Aws::Client::ClientConfiguration &clientConfig) {
    Aws::S3::S3Client s3_client(clientConfig);

    Aws::S3::Model::PutObjectRequest request;
    request.SetBucket(bucketName);
    request.SetKey(fileName);

    std::shared_ptr<Aws::IOStream> inputData =
            Aws::MakeShared<Aws::FStream>("SampleAllocationTag",
                                          filePath.c_str(),
                                          std::ios_base::in |
 std::ios_base::binary);

    if (!*inputData) {
        std::cerr << "Error unable to read file " << filePath << std::endl;
        return false;
    }

    request.SetBody(inputData);

    Aws::S3::Model::PutObjectOutcome outcome =
            s3_client.PutObject(request);

    if (!outcome.IsSuccess()) {
        std::cerr << "Error: PutObject: " <<
                  outcome.GetError().GetMessage() << std::endl;
    }
    else {
        std::cout << "Added object '" << filePath << "' to bucket '"
                  << bucketName << "'." << std::endl;
    }

    return outcome.IsSuccess();
}

//! Routine which deletes all objects in an S3 bucket.
/*!
 \\sa deleteAllObjectsInS3Bucket()
 \param bucketName: The S3 bucket name.
 \param clientConfig: AWS client configuration.
 \return bool: Successful completion.
 */
bool AwsDoc::Glue::deleteAllObjectsInS3Bucket(const Aws::String &bucketName,
                                              const
 Aws::Client::ClientConfiguration &clientConfig) {
    Aws::S3::S3Client client(clientConfig);
    Aws::S3::Model::ListObjectsV2Request listObjectsRequest;
    listObjectsRequest.SetBucket(bucketName);

    Aws::String continuationToken; // Used for pagination.
    bool result = true;
    do {
        if (!continuationToken.empty()) {
            listObjectsRequest.SetContinuationToken(continuationToken);
        }

        Aws::S3::Model::ListObjectsV2Outcome listObjectsOutcome =
client.ListObjectsV2(
                listObjectsRequest);

        if (listObjectsOutcome.IsSuccess()) {
            const std::vector<Aws::S3::Model::Object> &objects =
listObjectsOutcome.GetResult().GetContents();
            if (!objects.empty()) {
                Aws::S3::Model::DeleteObjectsRequest deleteObjectsRequest;
                deleteObjectsRequest.SetBucket(bucketName);

                std::vector<Aws::S3::Model::ObjectIdentifier> objectIdentifiers;
                for (const Aws::S3::Model::Object &object: objects) {
                    objectIdentifiers.push_back(
                            Aws::S3::Model::ObjectIdentifier().WithKey(
                                    object.GetKey()));
                }
                Aws::S3::Model::Delete objectsDelete;
                objectsDelete.SetObjects(objectIdentifiers);
                objectsDelete.SetQuiet(true);
                deleteObjectsRequest.SetDelete(objectsDelete);

                Aws::S3::Model::DeleteObjectsOutcome deleteObjectsOutcome =
                        client.DeleteObjects(deleteObjectsRequest);

                if (!deleteObjectsOutcome.IsSuccess()) {
                    std::cerr << "Error deleting objects. " <<
                              deleteObjectsOutcome.GetError().GetMessage() <<
std::endl;
                    result = false;
                    break;
                }
                else {
                    std::cout << "Successfully deleted the objects." <<
std::endl;

                }
            }
            else {
                std::cout << "No objects to delete in '" << bucketName << "'."
```

```
                            << std::endl;
              }

              continuationToken =
 listObjectsOutcome.GetResult().GetNextContinuationToken();
          }
          else {
              std::cerr << "Error listing objects. "
                        << listObjectsOutcome.GetError().GetMessage() << std::endl;
              result = false;
              break;
          }
    } while (!continuationToken.empty());

    return result;
}

//! Routine which retrieves an object from an S3 bucket.
/*!
 \\sa getObjectFromBucket()
 \param bucketName: The S3 bucket name.
 \param objectKey: The object's name.
 \param objectStream: A stream to receive the retrieved data.
 \param clientConfig: AWS client configuration.
 \return bool: Successful completion.
 */
bool AwsDoc::Glue::getObjectFromBucket(const Aws::String &bucketName,
                                       const Aws::String &objectKey,
                                       std::ostream &objectStream,
                                       const Aws::Client::ClientConfiguration
 &clientConfig) {
    Aws::S3::S3Client client(clientConfig);
    Aws::S3::Model::GetObjectRequest request;
    request.SetBucket(bucketName);
    request.SetKey(objectKey);

    Aws::S3::Model::GetObjectOutcome outcome = client.GetObject(request);


    if (outcome.IsSuccess()) {
        std::cout << "Successfully retrieved '" << objectKey << "'." <<
 std::endl;
        auto &body = outcome.GetResult().GetBody();
        objectStream << body.rdbuf();
    }
    else {
        std::cerr << "Error retrieving object. " <<
 outcome.GetError().GetMessage()
                  << std::endl;
    }

    return outcome.IsSuccess();
}
```

- For API details, see the following topics in *AWS SDK for C++ API Reference*.
  - CreateCrawler
  - CreateJob
  - DeleteCrawler
  - DeleteDatabase
  - DeleteJob
  - DeleteTable
  - GetCrawler
  - GetDatabase
  - GetDatabases
  - GetJob
  - GetJobRun
  - GetJobRuns
  - GetTables
  - ListJobs
  - StartCrawler
  - StartJobRun

Java

SDK for Java 2.x

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
/**
 * Before running this Java V2 code example, set up your development
 * environment, including your credentials.
 * <p>
 * For more information, see the following documentation topic:
 *
 * https://docs.aws.amazon.com/sdk-for-java/latest/developer-guide/get-
started.html
 *
 * To set up the resources, see this documentation topic:
 *
 * https://docs.aws.amazon.com/glue/latest/ug/tutorial-add-crawler.html
 *
 * This example performs the following tasks:
 *
 * 1. Create a database.
 * 2. Create a crawler.
 * 3. Get a crawler.
 * 4. Start a crawler.
 * 5. Get a database.
 * 6. Get tables.
 * 7. Create a job.
 * 8. Start a job run.
 * 9. List all jobs.
 * 10. Get job runs.
 * 11. Delete a job.
 * 12. Delete a database.
 * 13. Delete a crawler.
 */

public class GlueScenario {
    public static final String DASHES = new String(new char[80]).replace("\0",
"-");

    public static void main(String[] args) throws InterruptedException {
        final String usage = """

            Usage:
                <iam> <s3Path> <cron> <dbName> <crawlerName> <jobName>
<scriptLocation> <locationUri> <bucketNameSc>\s

            Where:
                iam - The ARN of the IAM role that has AWS Glue and S3
permissions.\s
                s3Path - The Amazon Simple Storage Service (Amazon S3) target
that contains data (for example, s3://<bucket name>/read).
                cron - A cron expression used to specify the schedule  (i.e.,
cron(15 12 * * ? *).
                dbName - The database name.\s
                crawlerName - The name of the crawler.\s
                jobName - The name you assign to this job definition.
                scriptLocation - The Amazon S3 path to a script that runs a job.
                locationUri - The location of the database (you can find this
file in resources folder).
                bucketNameSc - The Amazon S3 bucket name used when creating a job
                """;

        if (args.length != 9) {
            System.out.println(usage);
            return;
        }
        Scanner scanner = new Scanner(System.in);
        String iam = args[0];
        String s3Path = args[1];
        String cron = args[2];
        String dbName = args[3];
        String crawlerName = args[4];
        String jobName = args[5];
        String scriptLocation = args[6];
        String locationUri = args[7];
        String bucketNameSc = args[8];

        Region region = Region.US_EAST_1;
        GlueClient glueClient = GlueClient.builder()
            .region(region)
            .build();
        System.out.println(DASHES);
        System.out.println("Welcome to the AWS Glue scenario.");
        System.out.println("""
            AWS Glue is a fully managed extract, transform, and load (ETL)
 service provided by Amazon
            Web Services (AWS). It is designed to simplify the process of
 building, running, and maintaining
            ETL pipelines, which are essential for data integration and data
 warehousing tasks.

            One of the key features of AWS Glue is its ability to automatically
 discover and catalog data
            stored in various sources, such as Amazon S3, Amazon RDS, Amazon
 Redshift, and other databases.
            This cataloging process creates a central metadata repository, known
 as the AWS Glue Data Catalog,
            which provides a unified view of an organization's data assets. This
 metadata can then be used to
            create ETL jobs, which can be scheduled and run on-demand or on a
 regular basis.

            Lets get started.

            """);
        waitForInputToContinue(scanner);
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("1. Create a database.");
        try {
            createDatabase(glueClient, dbName, locationUri);
        } catch (GlueException e) {
            if (e.awsErrorDetails().errorMessage().equals("Database already
 exists.")) {
                System.out.println("Database " + dbName + " already exists.
 Skipping creation.");
            } else {
                System.err.println(e.awsErrorDetails().errorMessage());
                return;
            }
        }

        waitForInputToContinue(scanner);
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("2. Create a crawler.");
        try {
            createGlueCrawler(glueClient, iam, s3Path, cron, dbName,
 crawlerName);
        } catch (GlueException e) {
            if (e.awsErrorDetails().errorMessage().contains("already exists")) {
                System.out.println("Crawler " + crawlerName + " already exists.
 Skipping creation.");
            } else {
                System.err.println(e.awsErrorDetails().errorMessage());
                System.exit(1);
            }
        }
        waitForInputToContinue(scanner);
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("3. Get a crawler.");
        try {
            getSpecificCrawler(glueClient, crawlerName);
        } catch (GlueException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            return;
        }
        waitForInputToContinue(scanner);
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("4. Start a crawler.");
        try {
            startSpecificCrawler(glueClient, crawlerName);
        } catch (GlueException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            return;
        }
        waitForInputToContinue(scanner);
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("5. Get a database.");
        try {
                getSpecificDatabase(glueClient, dbName);
            } catch (GlueException e) {
                System.err.println(e.awsErrorDetails().errorMessage());
                return;
            }
            waitForInputToContinue(scanner);
            System.out.println(DASHES);

            System.out.println(DASHES);
            System.out.println("*** Wait 5 min for the tables to become available");
            TimeUnit.MINUTES.sleep(5);
            System.out.println("6. Get tables.");
            String myTableName;
            try {
                myTableName = getGlueTables(glueClient, dbName);
            } catch (GlueException e) {
                System.err.println(e.awsErrorDetails().errorMessage());
                return;
            }
            waitForInputToContinue(scanner);
            System.out.println(DASHES);

            System.out.println(DASHES);
            System.out.println("7. Create a job.");
            try {
                createJob(glueClient, jobName, iam, scriptLocation);
            } catch (GlueException e) {
                System.err.println(e.awsErrorDetails().errorMessage());
                return;
            }
            waitForInputToContinue(scanner);
            System.out.println(DASHES);

            System.out.println(DASHES);
            System.out.println("8. Start a Job run.");
            try {
                startJob(glueClient, jobName, dbName, myTableName, bucketNameSc);
            } catch (GlueException e) {
                System.err.println(e.awsErrorDetails().errorMessage());
                return;
            }
            waitForInputToContinue(scanner);
            System.out.println(DASHES);
```

```
        System.out.println(DASHES);
        System.out.println("9. List all jobs.");
        try {
            getAllJobs(glueClient);
        } catch (GlueException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            return;
        }
        waitForInputToContinue(scanner);
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("10. Get job runs.");
        try {
            getJobRuns(glueClient, jobName);
        } catch (GlueException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            return;
        }
        waitForInputToContinue(scanner);
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("11. Delete a job.");
        try {
            deleteJob(glueClient, jobName);
        } catch (GlueException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            return;
        }
        System.out.println("*** Wait 5 MIN for the " + crawlerName + " to stop");
        TimeUnit.MINUTES.sleep(5);
        waitForInputToContinue(scanner);
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("12. Delete a database.");
        try {
            deleteDatabase(glueClient, dbName);
        } catch (GlueException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            return;
        }
        waitForInputToContinue(scanner);
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("Delete a crawler.");
        try {
            deleteSpecificCrawler(glueClient, crawlerName);
        } catch (GlueException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            return;
        }
        waitForInputToContinue(scanner);
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("Successfully completed the AWS Glue Scenario");
        System.out.println(DASHES);
    }


    /**
     * Creates a Glue database with the specified name and location URI.
     *
     * @param glueClient  The Glue client to use for the database creation.
     * @param dbName      The name of the database to create.
     * @param locationUri The location URI for the database.
     */
    public static void createDatabase(GlueClient glueClient, String dbName,
 String locationUri) {
        try {
            DatabaseInput input = DatabaseInput.builder()
                .description("Built with the AWS SDK for Java V2")
                .name(dbName)
                .locationUri(locationUri)
                .build();

            CreateDatabaseRequest request = CreateDatabaseRequest.builder()
                .databaseInput(input)
                .build();

            glueClient.createDatabase(request);
            System.out.println(dbName + " was successfully created");

        } catch (GlueException e) {
            throw e;
        }
    }



    /**
     * Creates a new AWS Glue crawler using the AWS Glue Java API.
     *
     * @param glueClient  the AWS Glue client used to interact with the AWS Glue
 service
     * @param iam         the IAM role that the crawler will use to access the
 data source
     * @param s3Path      the S3 path that the crawler will scan for data
     * @param cron        the cron expression that defines the crawler's schedule
     * @param dbName      the name of the AWS Glue database where the crawler
 will store the metadata
     * @param crawlerName the name of the crawler to be created
     */
    public static void createGlueCrawler(GlueClient glueClient,
                                         String iam,
                                         String s3Path,
                                         String cron,
                                         String dbName,
                                         String crawlerName) {

        try {
            S3Target s3Target = S3Target.builder()
                .path(s3Path)
                .build();

            List<S3Target> targetList = new ArrayList<>();
            targetList.add(s3Target);
            CrawlerTargets targets = CrawlerTargets.builder()
                .s3Targets(targetList)
                .build();

            CreateCrawlerRequest crawlerRequest = CreateCrawlerRequest.builder()
                .databaseName(dbName)
                .name(crawlerName)
                .description("Created by the AWS Glue Java API")
                .targets(targets)
                .role(iam)
                .schedule(cron)
                .build();
            glueClient.createCrawler(crawlerRequest);
            System.out.println(crawlerName + " was successfully created");

        } catch (GlueException e) {
            throw e;
        }
    }

    /**
     * Retrieves a specific crawler from the AWS Glue service and waits for it to
 be in the "READY" state.
     *
     * @param glueClient  the AWS Glue client used to interact with the Glue
 service
     * @param crawlerName the name of the crawler to be retrieved
     */
    public static void getSpecificCrawler(GlueClient glueClient, String
 crawlerName) throws InterruptedException {
        try {
            GetCrawlerRequest crawlerRequest = GetCrawlerRequest.builder()
                .name(crawlerName)
                .build();

            boolean ready = false;
            while (!ready) {
                GetCrawlerResponse response =
 glueClient.getCrawler(crawlerRequest);
                String status = response.crawler().stateAsString();
                if (status.compareTo("READY") == 0) {
                    ready = true;
                }
                Thread.sleep(3000);
            }

            System.out.println("The crawler is now ready");

        } catch (GlueException | InterruptedException e) {
            throw e;
        }
    }

    /**
     * Starts a specific AWS Glue crawler.
     *
     * @param glueClient  the AWS Glue client to use for the crawler operation
     * @param crawlerName the name of the crawler to start
     * @throws GlueException if there is an error starting the crawler
     */
    public static void startSpecificCrawler(GlueClient glueClient, String
 crawlerName) {
        try {
            StartCrawlerRequest crawlerRequest = StartCrawlerRequest.builder()
                .name(crawlerName)
                .build();

            glueClient.startCrawler(crawlerRequest);
            System.out.println(crawlerName + " was successfully started!");

        } catch (GlueException e) {
            throw e;
        }
    }

    /**
     * Retrieves the specific database from the AWS Glue service.
     *
     * @param glueClient   an instance of the AWS Glue client used to interact
 with the service
     * @param databaseName the name of the database to retrieve
     * @throws GlueException if there is an error retrieving the database from
 the AWS Glue service
     */
    public static void getSpecificDatabase(GlueClient glueClient, String
 databaseName) {
        try {
            GetDatabaseRequest databasesRequest = GetDatabaseRequest.builder()
                .name(databaseName)
                .build();

            GetDatabaseResponse response =
 glueClient.getDatabase(databasesRequest);
            Instant createDate = response.database().createTime();

            // Convert the Instant to readable date.
            DateTimeFormatter formatter =
 DateTimeFormatter.ofLocalizedDateTime(FormatStyle.SHORT)
                .withLocale(Locale.US)
                .withZone(ZoneId.systemDefault());
            formatter.format(createDate);
            System.out.println("The create date of the database is " +
createDate);

        } catch (GlueException e) {
            throw e;
        }
    }



    /**
     * Retrieves the names of the tables in the specified Glue database.
     *
     * @param glueClient the Glue client to use for the operation
     * @param dbName     the name of the Glue database to retrieve the table
names from
     * @return the name of the first table retrieved, or an empty string if no
tables were found
     */
    public static String getGlueTables(GlueClient glueClient, String dbName) {
        String myTableName = "";
        try {
            GetTablesRequest tableRequest = GetTablesRequest.builder()
                .databaseName(dbName)
                .build();

            GetTablesResponse response = glueClient.getTables(tableRequest);
            List<Table> tables = response.tableList();
            if (tables.isEmpty()) {
                System.out.println("No tables were returned");
            } else {
                for (Table table : tables) {
                    myTableName = table.name();
                    System.out.println("Table name is: " + myTableName);
                }
            }

        } catch (GlueException e) {
            throw e;
        }
        return myTableName;
    }
    /**
     * Starts a job run in AWS Glue.
     *
     * @param glueClient    the AWS Glue client to use for the job run
     * @param jobName       the name of the Glue job to run
     * @param inputDatabase the name of the input database
     * @param inputTable    the name of the input table
     * @param outBucket     the URL of the output S3 bucket
     * @throws GlueException if there is an error starting the job run
     */
    public static void startJob(GlueClient glueClient, String jobName, String
 inputDatabase, String inputTable,
                                String outBucket) {
        try {
            Map<String, String> myMap = new HashMap<>();
            myMap.put("--input_database", inputDatabase);
            myMap.put("--input_table", inputTable);
            myMap.put("--output_bucket_url", outBucket);

            StartJobRunRequest runRequest = StartJobRunRequest.builder()
                .workerType(WorkerType.G_1_X)
                .numberOfWorkers(10)
                .arguments(myMap)
                .jobName(jobName)
                .build();

            StartJobRunResponse response = glueClient.startJobRun(runRequest);
            System.out.println("The request Id of the job is " +
 response.responseMetadata().requestId());

        } catch (GlueException e) {
            throw e;
        }
    }


    /**
     * Creates a new AWS Glue job.
     *
     * @param glueClient     the AWS Glue client to use for the operation
     * @param jobName        the name of the job to create
     * @param iam            the IAM role to associate with the job
     * @param scriptLocation the location of the script to be used by the job
     * @throws GlueException if there is an error creating the job
     */
    public static void createJob(GlueClient glueClient, String jobName, String
iam, String scriptLocation) {
        try {
            JobCommand command = JobCommand.builder()
                .pythonVersion("3")
                .name("glueetl")
                .scriptLocation(scriptLocation)
                .build();

            CreateJobRequest jobRequest = CreateJobRequest.builder()
                .description("A Job created by using the AWS SDK for Java V2")
                .glueVersion("2.0")
                .workerType(WorkerType.G_1_X)
                .numberOfWorkers(10)
                .name(jobName)
                .role(iam)
                .command(command)
                .build();

            glueClient.createJob(jobRequest);
            System.out.println(jobName + " was successfully created.");

        } catch (GlueException e) {
            throw e;
        }
    }


    /**
     * Retrieves and prints information about all the jobs in the Glue data
catalog.
     *
     * @param glueClient the Glue client used to interact with the AWS Glue
service
     */
    public static void getAllJobs(GlueClient glueClient) {
        try {
            GetJobsRequest jobsRequest = GetJobsRequest.builder()
                .maxResults(10)
                .build();

            GetJobsResponse jobsResponse = glueClient.getJobs(jobsRequest);
            List<Job> jobs = jobsResponse.jobs();
            for (Job job : jobs) {
                System.out.println("Job name is : " + job.name());
                System.out.println("The job worker type is : " +
 job.workerType().name());
            }

        } catch (GlueException e) {
            throw e;
        }
    }

    /**
     * Retrieves the job runs for a given Glue job and prints the status of the
 job runs.
     *
     * @param glueClient the Glue client used to make API calls
     * @param jobName    the name of the Glue job to retrieve the job runs for
     */
    public static void getJobRuns(GlueClient glueClient, String jobName) {
        try {
            GetJobRunsRequest runsRequest = GetJobRunsRequest.builder()
                .jobName(jobName)
                .maxResults(20)
                .build();

            boolean jobDone = false;
            while (!jobDone) {
                GetJobRunsResponse response = glueClient.getJobRuns(runsRequest);
                List<JobRun> jobRuns = response.jobRuns();
                for (JobRun jobRun : jobRuns) {
                    String jobState = jobRun.jobRunState().name();
                    if (jobState.compareTo("SUCCEEDED") == 0) {
                        System.out.println(jobName + " has succeeded");
                        jobDone = true;

                    } else if (jobState.compareTo("STOPPED") == 0) {
                        System.out.println("Job run has stopped");
                        jobDone = true;

                    } else if (jobState.compareTo("FAILED") == 0) {
                        System.out.println("Job run has failed");
                        jobDone = true;
                    } else if (jobState.compareTo("TIMEOUT") == 0) {
                        System.out.println("Job run has timed out");
                        jobDone = true;

                    } else {
                        System.out.println("*** Job run state is " +
jobRun.jobRunState().name());
                        System.out.println("Job run Id is " + jobRun.id());
                        System.out.println("The Glue version is " +
jobRun.glueVersion());
                    }
                    TimeUnit.SECONDS.sleep(5);
                }
            }

        } catch (GlueException e) {
            throw e;
        } catch (InterruptedException e) {
            throw new RuntimeException(e);
        }
    }


    /**
     * Deletes a Glue job.
     *
     * @param glueClient the Glue client to use for the operation
     * @param jobName    the name of the job to be deleted
     * @throws GlueException if there is an error deleting the job
     */
    public static void deleteJob(GlueClient glueClient, String jobName) {
        try {
            DeleteJobRequest jobRequest = DeleteJobRequest.builder()
                .jobName(jobName)
                .build();

            glueClient.deleteJob(jobRequest);
            System.out.println(jobName + " was successfully deleted");

        } catch (GlueException e) {
            throw e;
        }
    }
    /**
     * Deletes a AWS Glue Database.
     *
     * @param glueClient   An instance of the AWS Glue client used to interact
 with the AWS Glue service.
     * @param databaseName The name of the database to be deleted.
     * @throws GlueException If an error occurs while deleting the database.
     */
    public static void deleteDatabase(GlueClient glueClient, String databaseName)
 {
        try {
            DeleteDatabaseRequest request = DeleteDatabaseRequest.builder()
                .name(databaseName)
                .build();

            glueClient.deleteDatabase(request);
            System.out.println(databaseName + " was successfully deleted");

        } catch (GlueException e) {
            throw e;
        }
    }



    /**
     * Deletes a specific AWS Glue crawler.
     *
     * @param glueClient  the AWS Glue client object
     * @param crawlerName the name of the crawler to be deleted
     * @throws GlueException if an error occurs during the deletion process
     */
    public static void deleteSpecificCrawler(GlueClient glueClient, String
 crawlerName) {
        try {
            DeleteCrawlerRequest deleteCrawlerRequest =
 DeleteCrawlerRequest.builder()
                .name(crawlerName)
                .build();

            glueClient.deleteCrawler(deleteCrawlerRequest);
            System.out.println(crawlerName + " was deleted");

        } catch (GlueException e) {
            throw e;
        }
    }

    private static void waitForInputToContinue(Scanner scanner) {
        while (true) {
            System.out.println("");
            System.out.println("Enter 'c' followed by <ENTER> to continue:");
            String input = scanner.nextLine();

            if (input.trim().equalsIgnoreCase("c")) {
                System.out.println("Continuing with the program...");
                System.out.println("");
                break;
            } else {
                // Handle invalid input.
                System.out.println("Invalid input. Please try again.");
            }
        }
    }
}
```

- For API details, see the following topics in *AWS SDK for Java 2.x API Reference*.
  - CreateCrawler
  - CreateJob
  - DeleteCrawler
  - DeleteDatabase
  - DeleteJob
  - DeleteTable
  - GetCrawler
  - GetDatabase
  - GetDatabases
  - GetJob
  - GetJobRun
  - GetJobRuns
  - GetTables
  - ListJobs

- StartCrawler
- StartJobRun

JavaScript

SDK for JavaScript (v3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

Create and run a crawler that crawls a public Amazon Simple Storage Service (Amazon S3) bucket and generates a metadata database that describes the CSV-formatted data it finds.

```
const createCrawler = (name, role, dbName, tablePrefix, s3TargetPath) => {
  const client = new GlueClient({});

  const command = new CreateCrawlerCommand({
    Name: name,
    Role: role,
    DatabaseName: dbName,
    TablePrefix: tablePrefix,
    Targets: {
      S3Targets: [{ Path: s3TargetPath }],
    },
  });

  return client.send(command);
};

const getCrawler = (name) => {
  const client = new GlueClient({});

  const command = new GetCrawlerCommand({
    Name: name,
  });

  return client.send(command);
};
```

```
const startCrawler = (name) => {
  const client = new GlueClient({});

  const command = new StartCrawlerCommand({
    Name: name,
  });

  return client.send(command);
};

const crawlerExists = async ({ getCrawler }, crawlerName) => {
  try {
    await getCrawler(crawlerName);
    return true;
  } catch {
    return false;
  }
};

/**
 * @param {{ createCrawler: import('../../../actions/create-
crawler.js').createCrawler}} actions
 */
const makeCreateCrawlerStep = (actions) => async (context) => {
  if (await crawlerExists(actions, process.env.CRAWLER_NAME)) {
    log("Crawler already exists. Skipping creation.");
  } else {
    await actions.createCrawler(
      process.env.CRAWLER_NAME,
      process.env.ROLE_NAME,
      process.env.DATABASE_NAME,
      process.env.TABLE_PREFIX,
      process.env.S3_TARGET_PATH,
    );

    log("Crawler created successfully.", { type: "success" });
  }

  return { ...context };
};

/**
```

```
 * @param {(name: string) => Promise<import('@aws-sdk/client-
glue').GetCrawlerCommandOutput>} getCrawler
 * @param {string} crawlerName
 */
const waitForCrawler = async (getCrawler, crawlerName) => {
  const waitTimeInSeconds = 30;
  const { Crawler } = await getCrawler(crawlerName);

  if (!Crawler) {
    throw new Error(`Crawler with name ${crawlerName} not found.`);
  }

  if (Crawler.State === "READY") {
    return;
  }

  log(`Crawler is ${Crawler.State}. Waiting ${waitTimeInSeconds} seconds...`);
  await wait(waitTimeInSeconds);
  return waitForCrawler(getCrawler, crawlerName);
};

const makeStartCrawlerStep =
  ({ startCrawler, getCrawler }) =>
  async (context) => {
    log("Starting crawler.");
    await startCrawler(process.env.CRAWLER_NAME);
    log("Crawler started.", { type: "success" });

    log("Waiting for crawler to finish running. This can take a while.");
    await waitForCrawler(getCrawler, process.env.CRAWLER_NAME);
    log("Crawler ready.", { type: "success" });

    return { ...context };
  };
```

List information about databases and tables in your AWS Glue Data Catalog.

```
const getDatabase = (name) => {
  const client = new GlueClient({});

  const command = new GetDatabaseCommand({
    Name: name,
  });

  return client.send(command);
};

const getTables = (databaseName) => {
  const client = new GlueClient({});

  const command = new GetTablesCommand({
    DatabaseName: databaseName,
  });

  return client.send(command);
};

const makeGetDatabaseStep =
  ({ getDatabase }) =>
  async (context) => {
    const {
      Database: { Name },
    } = await getDatabase(process.env.DATABASE_NAME);
    log(`Database: ${Name}`);
    return { ...context };
  };

/**
 * @param {{ getTables: () => Promise<import('@aws-sdk/client-
glue').GetTablesCommandOutput}} config
 */
const makeGetTablesStep =
  ({ getTables }) =>
  async (context) => {
    const { TableList } = await getTables(process.env.DATABASE_NAME);
    log("Tables:");
    log(TableList.map((table) => `  • ${table.Name}\n`));
    return { ...context };
  };
```

Create and run a job that extracts CSV data from the source Amazon S3 bucket, transforms it by removing and renaming fields, and loads JSON-formatted output into another Amazon S3 bucket.

```
const createJob = (name, role, scriptBucketName, scriptKey) => {
  const client = new GlueClient({});

  const command = new CreateJobCommand({
    Name: name,
    Role: role,
    Command: {
      Name: "glueetl",
      PythonVersion: "3",
      ScriptLocation: `s3://${scriptBucketName}/${scriptKey}`,
    },
    GlueVersion: "3.0",
  });

  return client.send(command);
};

const startJobRun = (jobName, dbName, tableName, bucketName) => {
  const client = new GlueClient({});

  const command = new StartJobRunCommand({
    JobName: jobName,
    Arguments: {
      "--input_database": dbName,
      "--input_table": tableName,
      "--output_bucket_url": `s3://${bucketName}/`,
    },
  });

  return client.send(command);
};

const makeCreateJobStep =
  ({ createJob }) =>
  async (context) => {
    log("Creating Job.");
    await createJob(
      process.env.JOB_NAME,
      process.env.ROLE_NAME,
      process.env.BUCKET_NAME,
      process.env.PYTHON_SCRIPT_KEY,
    );
    log("Job created.", { type: "success" });
    return { ...context };
  };

/**
 * @param {(name: string, runId: string) => Promise<import('@aws-sdk/client-
glue').GetJobRunCommandOutput> }  getJobRun
 * @param {string} jobName
 * @param {string} jobRunId
 */
const waitForJobRun = async (getJobRun, jobName, jobRunId) => {
  const waitTimeInSeconds = 30;
  const { JobRun } = await getJobRun(jobName, jobRunId);

  if (!JobRun) {
    throw new Error(`Job run with id ${jobRunId} not found.`);
  }

  switch (JobRun.JobRunState) {
    case "FAILED":
    case "TIMEOUT":
    case "STOPPED":
    case "ERROR":
      throw new Error(
        `Job ${JobRun.JobRunState}. Error: ${JobRun.ErrorMessage}`,
      );
    case "SUCCEEDED":
      return;
    default:
      break;
  }

  log(
    `Job ${JobRun.JobRunState}. Waiting ${waitTimeInSeconds} more seconds...`,
  );
  await wait(waitTimeInSeconds);
  return waitForJobRun(getJobRun, jobName, jobRunId);
};

/**
 * @param {{ prompter: { prompt: () => Promise<{ shouldOpen: boolean }>} }}
 context
 */
const promptToOpen = async (context) => {
  const { shouldOpen } = await context.prompter.prompt({
    name: "shouldOpen",
    type: "confirm",
    message: "Open the output bucket in your browser?",
  });

  if (shouldOpen) {
    return open(
      `https://s3.console.aws.amazon.com/s3/buckets/${process.env.BUCKET_NAME} to
 view the output.`,
    );
  }
};

const makeStartJobRunStep =
  ({ startJobRun, getJobRun }) =>
  async (context) => {
    log("Starting job.");
    const { JobRunId } = await startJobRun(
      process.env.JOB_NAME,
      process.env.DATABASE_NAME,
      process.env.TABLE_NAME,
      process.env.BUCKET_NAME,
    );
    log("Job started.", { type: "success" });

    log("Waiting for job to finish running. This can take a while.");
    await waitForJobRun(getJobRun, process.env.JOB_NAME, JobRunId);
    log("Job run succeeded.", { type: "success" });

    await promptToOpen(context);

    return { ...context };
  };
```

List information about job runs and view some of the transformed data.

```
const getJobRuns = (jobName) => {
  const client = new GlueClient({});
  const command = new GetJobRunsCommand({
    JobName: jobName,
  });
  return client.send(command);
};

const getJobRun = (jobName, jobRunId) => {
  const client = new GlueClient({});
  const command = new GetJobRunCommand({
    JobName: jobName,
    RunId: jobRunId,
  });

  return client.send(command);
};

/**
 * @typedef {{ prompter: { prompt: () => Promise<{jobName: string}> } }} Context
 */

/**
 * @typedef {() => Promise<import('@aws-sdk/client-
glue').GetJobRunCommandOutput>} getJobRun
 */

/**
 * @typedef {() => Promise<import('@aws-sdk/client-
glue').GetJobRunsCommandOutput} getJobRuns
 */

/**
 *
 * @param {getJobRun} getJobRun
 * @param {string} jobName
 * @param {string} jobRunId
 */
const logJobRunDetails = async (getJobRun, jobName, jobRunId) => {
  const { JobRun } = await getJobRun(jobName, jobRunId);
  log(JobRun, { type: "object" });
};

/**
 *
 * @param {{getJobRuns: getJobRuns, getJobRun: getJobRun }} funcs
 */
const makePickJobRunStep =
  ({ getJobRuns, getJobRun }) =>
  async (/** @type { Context } */ context) => {
    if (context.selectedJobName) {
      const { JobRuns } = await getJobRuns(context.selectedJobName);

      const { jobRunId } = await context.prompter.prompt({
        name: "jobRunId",
        type: "list",
        message: "Select a job run to see details.",
        choices: JobRuns.map((run) => run.Id),
      });

      logJobRunDetails(getJobRun, context.selectedJobName, jobRunId);
    }

    return { ...context };
  };
```

Delete all resources created by the demo.

```
const deleteJob = (jobName) => {
  const client = new GlueClient({});

  const command = new DeleteJobCommand({
    JobName: jobName,
  });

  return client.send(command);
};

const deleteTable = (databaseName, tableName) => {
  const client = new GlueClient({});

  const command = new DeleteTableCommand({
    DatabaseName: databaseName,
    Name: tableName,
  });

  return client.send(command);
};

const deleteDatabase = (databaseName) => {
  const client = new GlueClient({});

  const command = new DeleteDatabaseCommand({
    Name: databaseName,
  });

  return client.send(command);
};

const deleteCrawler = (crawlerName) => {
  const client = new GlueClient({});

  const command = new DeleteCrawlerCommand({
    Name: crawlerName,
  });

  return client.send(command);
};

/**
 *
 * @param {import('../../../actions/delete-job.js').deleteJob} deleteJobFn
 * @param {string[]} jobNames
 * @param {{ prompter: { prompt: () => Promise<any> }}} context
 */
const handleDeleteJobs = async (deleteJobFn, jobNames, context) => {
  /**
   * @type {{ selectedJobNames: string[] }}
   */
  const { selectedJobNames } = await context.prompter.prompt({
    name: "selectedJobNames",
    type: "checkbox",
    message: "Let's clean up jobs. Select jobs to delete.",
    choices: jobNames,
  });

  if (selectedJobNames.length === 0) {
    log("No jobs selected.");
  } else {
    log("Deleting jobs.");
    await Promise.all(
      selectedJobNames.map((n) => deleteJobFn(n).catch(console.error)),
    );
    log("Jobs deleted.", { type: "success" });
  }
};

/**
 * @param {{
 *   listJobs: import('../../../actions/list-jobs.js').listJobs,
 *   deleteJob: import('../../../actions/delete-job.js').deleteJob
 * }} config
 */
const makeCleanUpJobsStep =
  ({ listJobs, deleteJob }) =>
  async (context) => {
    const { JobNames } = await listJobs();
    if (JobNames.length > 0) {
      await handleDeleteJobs(deleteJob, JobNames, context);
    }

    return { ...context };
  };

/**
 * @param {import('../../../actions/delete-table.js').deleteTable} deleteTable
 * @param {string} databaseName
 * @param {string[]} tableNames
 */
const deleteTables = (deleteTable, databaseName, tableNames) =>
  Promise.all(
    tableNames.map((tableName) =>
      deleteTable(databaseName, tableName).catch(console.error),
    ),
  );

/**
 * @param {{
 *  getTables: import('../../../actions/get-tables.js').getTables,
 *  deleteTable: import('../../../actions/delete-table.js').deleteTable
 * }} config
 */
const makeCleanUpTablesStep =
  ({ getTables, deleteTable }) =>
  /**
   * @param {{ prompter: { prompt: () => Promise<any>}}} context
   */
  async (context) => {
    const { TableList } = await getTables(process.env.DATABASE_NAME).catch(
      () => ({ TableList: null }),
    );

    if (TableList && TableList.length > 0) {
      /**
       * @type {{ tableNames: string[] }}
       */
      const { tableNames } = await context.prompter.prompt({
        name: "tableNames",
        type: "checkbox",
        message: "Let's clean up tables. Select tables to delete.",
        choices: TableList.map((t) => t.Name),
      });

      if (tableNames.length === 0) {
        log("No tables selected.");
      } else {
        log("Deleting tables.");
        await deleteTables(deleteTable, process.env.DATABASE_NAME, tableNames);
        log("Tables deleted.", { type: "success" });
      }
    }

    return { ...context };
  };

/**
 * @param {import('../../../actions/delete-database.js').deleteDatabase}
 deleteDatabase
 * @param {string[]} databaseNames
 */
const deleteDatabases = (deleteDatabase, databaseNames) =>
  Promise.all(
    databaseNames.map((dbName) => deleteDatabase(dbName).catch(console.error)),
  );

/**
 * @param {{
 *   getDatabases: import('../../../actions/get-databases.js').getDatabases
 *   deleteDatabase: import('../../../actions/delete-database.js').deleteDatabase
 * }} config
 */
const makeCleanUpDatabasesStep =
  ({ getDatabases, deleteDatabase }) =>
  /**
   * @param {{ prompter: { prompt: () => Promise<any>}} context
   */
  async (context) => {
    const { DatabaseList } = await getDatabases();

    if (DatabaseList.length > 0) {
      /** @type {{ dbNames: string[] }} */
      const { dbNames } = await context.prompter.prompt({
        name: "dbNames",
        type: "checkbox",
        message: "Let's clean up databases. Select databases to delete.",
        choices: DatabaseList.map((db) => db.Name),
      });

      if (dbNames.length === 0) {
        log("No databases selected.");
      } else {
        log("Deleting databases.");
        await deleteDatabases(deleteDatabase, dbNames);
        log("Databases deleted.", { type: "success" });
      }
    }

    return { ...context };
  };

const cleanUpCrawlerStep = async (context) => {
  log("Deleting crawler.");

  try {
    await deleteCrawler(process.env.CRAWLER_NAME);
    log("Crawler deleted.", { type: "success" });
  } catch (err) {
    if (err.name === "EntityNotFoundException") {
      log("Crawler is already deleted.");
    } else {
      throw err;
    }
  }

  return { ...context };
};
```

- For API details, see the following topics in *AWS SDK for JavaScript API Reference*.
  - CreateCrawler
  - CreateJob
  - DeleteCrawler
  - DeleteDatabase
  - DeleteJob
  - DeleteTable
  - GetCrawler
  - GetDatabase
  - GetDatabases
  - GetJob
  - GetJobRun
  - GetJobRuns
  - GetTables
  - ListJobs
  - StartCrawler
  - StartJobRun

Kotlin

SDK for Kotlin

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
suspend fun main(args: Array<String>) {
    val usage = """
        Usage:
            <iam> <s3Path> <cron> <dbName> <crawlerName> <jobName>
 <scriptLocation> <locationUri>
```

```
       Where:
           iam - The Amazon Resource Name (ARN) of the AWS Identity and Access
Management (IAM) role that has AWS Glue and Amazon Simple Storage Service
(Amazon S3) permissions.
           s3Path - The Amazon Simple Storage Service (Amazon S3) target that
contains data (for example, CSV data).
           cron - A cron expression used to specify the schedule (for example,
cron(15 12 * * ? *).
           dbName - The database name.
           crawlerName - The name of the crawler.
           jobName - The name you assign to this job definition.
           scriptLocation - Specifies the Amazon S3 path to a script that runs a
job.
           locationUri - Specifies the location of the database
       """

   if (args.size != 8) {
       println(usage)
       exitProcess(1)
   }

   val iam = args[0]
   val s3Path = args[1]
   val cron = args[2]
   val dbName = args[3]
   val crawlerName = args[4]
   val jobName = args[5]
   val scriptLocation = args[6]
   val locationUri = args[7]

   println("About to start the AWS Glue Scenario")
   createDatabase(dbName, locationUri)
   createCrawler(iam, s3Path, cron, dbName, crawlerName)
   getCrawler(crawlerName)
   startCrawler(crawlerName)
   getDatabase(dbName)
   getGlueTables(dbName)
   createJob(jobName, iam, scriptLocation)
   startJob(jobName)
   getJobs()
   getJobRuns(jobName)
   deleteJob(jobName)
   println("*** Wait for 5 MIN so the $crawlerName is ready to be deleted")
```

```
    TimeUnit.MINUTES.sleep(5)
    deleteMyDatabase(dbName)
    deleteCrawler(crawlerName)
}

suspend fun createDatabase(
    dbName: String?,
    locationUriVal: String?,
) {
    val input =
        DatabaseInput {
            description = "Built with the AWS SDK for Kotlin"
            name = dbName
            locationUri = locationUriVal
        }

    val request =
        CreateDatabaseRequest {
            databaseInput = input
        }

    GlueClient.fromEnvironment { region = "us-east-1" }.use { glueClient ->
        glueClient.createDatabase(request)
        println("The database was successfully created")
    }
}

suspend fun createCrawler(
    iam: String?,
    s3Path: String?,
    cron: String?,
    dbName: String?,
    crawlerName: String,
) {
    val s3Target =
        S3Target {
            path = s3Path
        }

    val targetList = ArrayList<S3Target>()
    targetList.add(s3Target)

    val targetOb =
        CrawlerTargets {
            s3Targets = targetList
        }

    val crawlerRequest =
        CreateCrawlerRequest {
            databaseName = dbName
            name = crawlerName
            description = "Created by the AWS Glue Java API"
            targets = targetOb
            role = iam
            schedule = cron
        }

    GlueClient.fromEnvironment { region = "us-east-1" }.use { glueClient ->
        glueClient.createCrawler(crawlerRequest)
        println("$crawlerName was successfully created")
    }
}

suspend fun getCrawler(crawlerName: String?) {
    val request =
        GetCrawlerRequest {
            name = crawlerName
        }

    GlueClient.fromEnvironment { region = "us-east-1" }.use { glueClient ->
        val response = glueClient.getCrawler(request)
        val role = response.crawler?.role
        println("The role associated with this crawler is $role")
    }
}

suspend fun startCrawler(crawlerName: String) {
    val crawlerRequest =
        StartCrawlerRequest {
            name = crawlerName
        }

    GlueClient.fromEnvironment { region = "us-east-1" }.use { glueClient ->
        glueClient.startCrawler(crawlerRequest)
        println("$crawlerName was successfully started.")
    }
}
```

```
suspend fun getDatabase(databaseName: String?) {
    val request =
        GetDatabaseRequest {
            name = databaseName
        }

    GlueClient.fromEnvironment { region = "us-east-1" }.use { glueClient ->
        val response = glueClient.getDatabase(request)
        val dbDesc = response.database?.description
        println("The database description is $dbDesc")
    }
}

suspend fun getGlueTables(dbName: String?) {
    val tableRequest =
        GetTablesRequest {
            databaseName = dbName
        }

    GlueClient.fromEnvironment { region = "us-east-1" }.use { glueClient ->
        val response = glueClient.getTables(tableRequest)
        response.tableList?.forEach { tableName ->
            println("Table name is ${tableName.name}")
        }
    }
}

suspend fun startJob(jobNameVal: String?) {
    val runRequest =
        StartJobRunRequest {
            workerType = WorkerType.G1X
            numberOfWorkers = 10
            jobName = jobNameVal
        }

    GlueClient.fromEnvironment { region = "us-east-1" }.use { glueClient ->
        val response = glueClient.startJobRun(runRequest)
        println("The job run Id is ${response.jobRunId}")
    }
}

suspend fun createJob(
    jobName: String,
    iam: String?,
    scriptLocationVal: String?,
) {
    val commandOb =
        JobCommand {
            pythonVersion = "3"
            name = "MyJob1"
            scriptLocation = scriptLocationVal
        }

    val jobRequest =
        CreateJobRequest {
            description = "A Job created by using the AWS SDK for Java V2"
            glueVersion = "2.0"
            workerType = WorkerType.G1X
            numberOfWorkers = 10
            name = jobName
            role = iam
            command = commandOb
        }

    GlueClient.fromEnvironment { region = "us-east-1" }.use { glueClient ->
        glueClient.createJob(jobRequest)
        println("$jobName was successfully created.")
    }
}

suspend fun getJobs() {
    val request =
        GetJobsRequest {
            maxResults = 10
        }

    GlueClient.fromEnvironment { region = "us-east-1" }.use { glueClient ->
        val response = glueClient.getJobs(request)
        response.jobs?.forEach { job ->
            println("Job name is ${job.name}")
        }
    }
}

suspend fun getJobRuns(jobNameVal: String?) {
    val request =
        GetJobRunsRequest {
            jobName = jobNameVal
        }

    GlueClient.fromEnvironment { region = "us-east-1" }.use { glueClient ->
        val response = glueClient.getJobRuns(request)
        response.jobRuns?.forEach { job ->
            println("Job name is ${job.jobName}")
        }
    }
}

suspend fun deleteJob(jobNameVal: String) {
    val jobRequest =
        DeleteJobRequest {
            jobName = jobNameVal
        }

    GlueClient.fromEnvironment { region = "us-east-1" }.use { glueClient ->
        glueClient.deleteJob(jobRequest)
        println("$jobNameVal was successfully deleted")
    }
}

suspend fun deleteMyDatabase(databaseName: String) {
    val request =
        DeleteDatabaseRequest {
            name = databaseName
        }

    GlueClient.fromEnvironment { region = "us-east-1" }.use { glueClient ->
        glueClient.deleteDatabase(request)
        println("$databaseName was successfully deleted")
    }
}

suspend fun deleteCrawler(crawlerName: String) {
    val request =
        DeleteCrawlerRequest {
            name = crawlerName
        }
    GlueClient.fromEnvironment { region = "us-east-1" }.use { glueClient ->
        glueClient.deleteCrawler(request)
        println("$crawlerName was deleted")
    }
}
```

- For API details, see the following topics in *AWS SDK for Kotlin API reference*.
  - CreateCrawler
  - CreateJob
  - DeleteCrawler
  - DeleteDatabase
  - DeleteJob
  - DeleteTable
  - GetCrawler
  - GetDatabase
  - GetDatabases
  - GetJob
  - GetJobRun
  - GetJobRuns
  - GetTables
  - ListJobs
  - StartCrawler
  - StartJobRun

PHP

SDK for PHP

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
namespace Glue;

use Aws\Glue\GlueClient;
use Aws\S3\S3Client;
use AwsUtilities\AWSServiceClass;
use GuzzleHttp\Psr7\Stream;
use Iam\IAMService;

class GettingStartedWithGlue
{
    public function run()
    {
        echo("\n");
        echo("--------------------------------------\n");
        print("Welcome to the AWS Glue getting started demo using PHP!\n");
        echo("--------------------------------------\n");

        $clientArgs = [
            'region' => 'us-west-2',
            'version' => 'latest',
            'profile' => 'default',
        ];
        $uniqid = uniqid();

        $glueClient = new GlueClient($clientArgs);
        $glueService = new GlueService($glueClient);
        $iamService = new IAMService();
        $crawlerName = "example-crawler-test-" . $uniqid;

        AWSServiceClass::$waitTime = 5;
        AWSServiceClass::$maxWaitAttempts = 20;

        $role = $iamService->getRole("AWSGlueServiceRole-DocExample");

        $databaseName = "doc-example-database-$uniqid";
        $path = 's3://crawler-public-us-east-1/flight/2016/csv';
        $glueService->createCrawler($crawlerName, $role['Role']['Arn'],
 $databaseName, $path);
        $glueService->startCrawler($crawlerName);

        echo "Waiting for crawler";
        do {
            $crawler = $glueService->getCrawler($crawlerName);
            echo ".";
            sleep(10);
        } while ($crawler['Crawler']['State'] != "READY");
        echo "\n";

        $database = $glueService->getDatabase($databaseName);
        echo "Found a database named " . $database['Database']['Name'] . "\n";

        //Upload job script
        $s3client = new S3Client($clientArgs);
        $bucketName = "test-glue-bucket-" . $uniqid;
        $s3client->createBucket([
            'Bucket' => $bucketName,
            'CreateBucketConfiguration' => ['LocationConstraint' => 'us-west-2'],
        ]);

        $s3client->putObject([
            'Bucket' => $bucketName,
            'Key' => 'run_job.py',
            'SourceFile' => __DIR__ . '/flight_etl_job_script.py'
        ]);
        $s3client->putObject([
            'Bucket' => $bucketName,
            'Key' => 'setup_scenario_getting_started.yaml',
            'SourceFile' => __DIR__ . '/setup_scenario_getting_started.yaml'
        ]);

        $tables = $glueService->getTables($databaseName);

        $jobName = 'test-job-' . $uniqid;
        $scriptLocation = "s3://$bucketName/run_job.py";
        $job = $glueService->createJob($jobName, $role['Role']['Arn'],
 $scriptLocation);

        $outputBucketUrl = "s3://$bucketName";
        $runId = $glueService->startJobRun($jobName, $databaseName, $tables,
 $outputBucketUrl)['JobRunId'];

        echo "waiting for job";
        do {
            $jobRun = $glueService->getJobRun($jobName, $runId);
            echo ".";
            sleep(10);
        } while (!array_intersect([$jobRun['JobRun']['JobRunState']],
 ['SUCCEEDED', 'STOPPED', 'FAILED', 'TIMEOUT']));
        echo "\n";

        $jobRuns = $glueService->getJobRuns($jobName);

        $objects = $s3client->listObjects([
            'Bucket' => $bucketName,
        ])['Contents'];

        foreach ($objects as $object) {
            echo $object['Key'] . "\n";
        }

        echo "Downloading " . $objects[1]['Key'] . "\n";
        /** @var Stream $downloadObject */
        $downloadObject = $s3client->getObject([
            'Bucket' => $bucketName,
            'Key' => $objects[1]['Key'],
        ])['Body']->getContents();
        echo "Here is the first 1000 characters in the object.";
        echo substr($downloadObject, 0, 1000);

        $jobs = $glueService->listJobs();
        echo "Current jobs:\n";
        foreach ($jobs['JobNames'] as $jobsName) {
            echo "{$jobsName}\n";
        }

        echo "Delete the job.\n";
        $glueClient->deleteJob([
            'JobName' => $job['Name'],
        ]);

        echo "Delete the tables.\n";
        foreach ($tables['TableList'] as $table) {
            $glueService->deleteTable($table['Name'], $databaseName);
        }

        echo "Delete the databases.\n";
        $glueClient->deleteDatabase([
            'Name' => $databaseName,
        ]);

        echo "Delete the crawler.\n";
        $glueClient->deleteCrawler([
            'Name' => $crawlerName,
        ]);

        $deleteObjects = $s3client->listObjectsV2([
            'Bucket' => $bucketName,
        ]);
        echo "Delete all objects in the bucket.\n";
        $deleteObjects = $s3client->deleteObjects([
            'Bucket' => $bucketName,
            'Delete' => [
                'Objects' => $deleteObjects['Contents'],
            ]
        ]);
        echo "Delete the bucket.\n";
        $s3client->deleteBucket(['Bucket' => $bucketName]);

        echo "This job was brought to you by the number $uniqid\n";
    }
}

namespace Glue;

use Aws\Glue\GlueClient;
use Aws\Result;

use function PHPUnit\Framework\isEmpty;

class GlueService extends \AwsUtilities\AWSServiceClass
{
    protected GlueClient $glueClient;

    public function __construct($glueClient)
    {
        $this->glueClient = $glueClient;
    }

    public function getCrawler($crawlerName)
    {
        return $this->customWaiter(function () use ($crawlerName) {
            return $this->glueClient->getCrawler([
                'Name' => $crawlerName,
            ]);
        });
    }

    public function createCrawler($crawlerName, $role, $databaseName, $path):
 Result
    {
        return $this->customWaiter(function () use ($crawlerName, $role,
 $databaseName, $path) {
            return $this->glueClient->createCrawler([
                'Name' => $crawlerName,
                'Role' => $role,
                'DatabaseName' => $databaseName,
                'Targets' => [
                    'S3Targets' =>
                        [[
                            'Path' => $path,
                        ]]
                ],
            ]);
        });
    }

    public function startCrawler($crawlerName): Result
    {
        return $this->glueClient->startCrawler([
            'Name' => $crawlerName,
        ]);
    }

    public function getDatabase(string $databaseName): Result
    {
        return $this->customWaiter(function () use ($databaseName) {
            return $this->glueClient->getDatabase([
                'Name' => $databaseName,
            ]);
        });
    }

    public function getTables($databaseName): Result
    {
        return $this->glueClient->getTables([
            'DatabaseName' => $databaseName,
        ]);
    }

    public function createJob($jobName, $role, $scriptLocation, $pythonVersion =
 '3', $glueVersion = '3.0'): Result
    {
        return $this->glueClient->createJob([
            'Name' => $jobName,
            'Role' => $role,
            'Command' => [
                'Name' => 'glueetl',
                'ScriptLocation' => $scriptLocation,
                'PythonVersion' => $pythonVersion,
            ],
            'GlueVersion' => $glueVersion,
        ]);
    }

    public function startJobRun($jobName, $databaseName, $tables,
$outputBucketUrl): Result
    {
        return $this->glueClient->startJobRun([
            'JobName' => $jobName,
            'Arguments' => [
                'input_database' => $databaseName,
                'input_table' => $tables['TableList'][0]['Name'],
                'output_bucket_url' => $outputBucketUrl,
                '--input_database' => $databaseName,
                '--input_table' => $tables['TableList'][0]['Name'],
                '--output_bucket_url' => $outputBucketUrl,
            ],
        ]);
    }

    public function listJobs($maxResults = null, $nextToken = null, $tags = []):
Result
    {
        $arguments = [];
        if ($maxResults) {
            $arguments['MaxResults'] = $maxResults;
        }
        if ($nextToken) {
            $arguments['NextToken'] = $nextToken;
        }
        if (!empty($tags)) {
            $arguments['Tags'] = $tags;
        }
        return $this->glueClient->listJobs($arguments);
    }

    public function getJobRuns($jobName, $maxResults = 0, $nextToken = ''):
Result
```

```
    {
        $arguments = ['JobName' => $jobName];
        if ($maxResults) {
            $arguments['MaxResults'] = $maxResults;
        }
        if ($nextToken) {
            $arguments['NextToken'] = $nextToken;
        }
        return $this->glueClient->getJobRuns($arguments);
    }

    public function getJobRun($jobName, $runId, $predecessorsIncluded = false):
 Result
    {
        return $this->glueClient->getJobRun([
            'JobName' => $jobName,
            'RunId' => $runId,
            'PredecessorsIncluded' => $predecessorsIncluded,
        ]);
    }

    public function deleteJob($jobName)
    {
        return $this->glueClient->deleteJob([
            'JobName' => $jobName,
        ]);
    }

    public function deleteTable($tableName, $databaseName)
    {
        return $this->glueClient->deleteTable([
            'DatabaseName' => $databaseName,
            'Name' => $tableName,
        ]);
    }

    public function deleteDatabase($databaseName)
    {
        return $this->glueClient->deleteDatabase([
            'Name' => $databaseName,
        ]);
    }

    public function deleteCrawler($crawlerName)
    {
        return $this->glueClient->deleteCrawler([
            'Name' => $crawlerName,
        ]);
    }
}
```

- For API details, see the following topics in *AWS SDK for PHP API Reference*.
  - CreateCrawler
  - CreateJob
  - DeleteCrawler
  - DeleteDatabase
  - DeleteJob
  - DeleteTable
  - GetCrawler
  - GetDatabase
  - GetDatabases
  - GetJob
  - GetJobRun
  - GetJobRuns
  - GetTables
  - ListJobs
  - StartCrawler
  - StartJobRun

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

Create a class that wraps AWS Glue functions used in the scenario.

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""

    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 Glue client.
        """
        self.glue_client = glue_client


    def get_crawler(self, name):
        """
        Gets information about a crawler.

        :param name: The name of the crawler to look up.
        :return: Data about the crawler.
        """
        crawler = None
        try:
            response = self.glue_client.get_crawler(Name=name)
            crawler = response["Crawler"]
        except ClientError as err:
            if err.response["Error"]["Code"] == "EntityNotFoundException":
                logger.info("Crawler %s doesn't exist.", name)
            else:
                logger.error(
                    "Couldn't get crawler %s. Here's why: %s: %s",
                    name,
                    err.response["Error"]["Code"],
                    err.response["Error"]["Message"],
                )
                raise
        return crawler


    def create_crawler(self, name, role_arn, db_name, db_prefix, s3_target):
        """
        Creates a crawler that can crawl the specified target and populate a
        database in your AWS Glue Data Catalog with metadata that describes the
 data
        in the target.
        :param name: The name of the crawler.
        :param role_arn: The Amazon Resource Name (ARN) of an AWS Identity and 
 Access
                         Management (IAM) role that grants permission to let AWS 
 Glue
                         access the resources it needs.
        :param db_name: The name to give the database that is created by the 
 crawler.
        :param db_prefix: The prefix to give any database tables that are created 
 by
                          the crawler.
        :param s3_target: The URL to an S3 bucket that contains data that is
                          the target of the crawler.
        """
        try:
            self.glue_client.create_crawler(
                Name=name,
                Role=role_arn,
                DatabaseName=db_name,
                TablePrefix=db_prefix,
                Targets={"S3Targets": [{"Path": s3_target}]},
            )
        except ClientError as err:
            logger.error(
                "Couldn't create crawler. Here's why: %s: %s",
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise


    def start_crawler(self, name):
        """
        Starts a crawler. The crawler crawls its configured target and creates
        metadata that describes the data it finds in the target data source.

        :param name: The name of the crawler to start.
        """
        try:
            self.glue_client.start_crawler(Name=name)
        except ClientError as err:
            logger.error(
                "Couldn't start crawler %s. Here's why: %s: %s",
                name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise


    def get_database(self, name):
        """
        Gets information about a database in your Data Catalog.

        :param name: The name of the database to look up.
        :return: Information about the database.
        """
        try:
            response = self.glue_client.get_database(Name=name)
        except ClientError as err:
            logger.error(
                "Couldn't get database %s. Here's why: %s: %s",
                name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise
        else:
            return response["Database"]


    def get_tables(self, db_name):
        """
        Gets a list of tables in a Data Catalog database.

        :param db_name: The name of the database to query.
        :return: The list of tables in the database.
        """
        try:
            response = self.glue_client.get_tables(DatabaseName=db_name)
        except ClientError as err:
            logger.error(
                "Couldn't get tables %s. Here's why: %s: %s",
                db_name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise
        else:
            return response["TableList"]


    def create_job(self, name, description, role_arn, script_location):
        """
        Creates a job definition for an extract, transform, and load (ETL) job
 that can
        be run by AWS Glue.

        :param name: The name of the job definition.
        :param description: The description of the job definition.
        :param role_arn: The ARN of an IAM role that grants AWS Glue the
 permissions
                         it requires to run the job.
        :param script_location: The Amazon S3 URL of a Python ETL script that is
 run as
                                part of the job. The script defines how the data
 is
                                transformed.
        """
        try:
            self.glue_client.create_job(
                Name=name,
                Description=description,
                Role=role_arn,
                Command={
                    "Name": "glueetl",
                    "ScriptLocation": script_location,
                    "PythonVersion": "3",
                },
                GlueVersion="3.0",
            )
        except ClientError as err:
            logger.error(
                "Couldn't create job %s. Here's why: %s: %s",
                name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise
    def start_job_run(self, name, input_database, input_table, 
output_bucket_name):
        """
        Starts a job run. A job run extracts data from the source, transforms it,
        and loads it to the output bucket.

        :param name: The name of the job definition.
        :param input_database: The name of the metadata database that contains 
tables
                               that describe the source data. This is typically 
created
                               by a crawler.
        :param input_table: The name of the table in the metadata database that
                            describes the source data.
        :param output_bucket_name: The S3 bucket where the output is written.
        :return: The ID of the job run.
        """
        try:
            # The custom Arguments that are passed to this function are used by 
the
            # Python ETL script to determine the location of input and output 
data.
            response = self.glue_client.start_job_run(
                JobName=name,
                Arguments={
                    "--input_database": input_database,
                    "--input_table": input_table,
                    "--output_bucket_url": f"s3://{output_bucket_name}/",
                },
            )
        except ClientError as err:
            logger.error(
                "Couldn't start job run %s. Here's why: %s: %s",
                name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise
        else:
            return response["JobRunId"]


    def list_jobs(self):
        """
        Lists the names of job definitions in your account.

        :return: The list of job definition names.
        """
        try:
            response = self.glue_client.list_jobs()
        except ClientError as err:
            logger.error(
                "Couldn't list jobs. Here's why: %s: %s",
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise
        else:
            return response["JobNames"]


    def get_job_runs(self, job_name):
        """
        Gets information about runs that have been performed for a specific job
        definition.

        :param job_name: The name of the job definition to look up.
        :return: The list of job runs.
        """
        try:
            response = self.glue_client.get_job_runs(JobName=job_name)
        except ClientError as err:
            logger.error(
                "Couldn't get job runs for %s. Here's why: %s: %s",
                job_name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise
        else:
            return response["JobRuns"]


    def get_job_run(self, name, run_id):
        """
        Gets information about a single job run.

        :param name: The name of the job definition for the run.
        :param run_id: The ID of the run.
        :return: Information about the run.
        """
        try:
            response = self.glue_client.get_job_run(JobName=name, RunId=run_id)
        except ClientError as err:
            logger.error(
                "Couldn't get job run %s/%s. Here's why: %s: %s",
                name,
                run_id,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise
        else:
            return response["JobRun"]


    def delete_job(self, job_name):
        """
        Deletes a job definition. This also deletes data about all runs that are
        associated with this job definition.

        :param job_name: The name of the job definition to delete.
        """
        try:
            self.glue_client.delete_job(JobName=job_name)
        except ClientError as err:
            logger.error(
                "Couldn't delete job %s. Here's why: %s: %s",
                job_name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise


    def delete_table(self, db_name, table_name):
        """
        Deletes a table from a metadata database.

        :param db_name: The name of the database that contains the table.
        :param table_name: The name of the table to delete.
        """
        try:
            self.glue_client.delete_table(DatabaseName=db_name, Name=table_name)
        except ClientError as err:
            logger.error(
                "Couldn't delete table %s. Here's why: %s: %s",
                table_name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise


    def delete_database(self, name):
        """
        Deletes a metadata database from your Data Catalog.

        :param name: The name of the database to delete.
        """
        try:
            self.glue_client.delete_database(Name=name)
        except ClientError as err:
            logger.error(
                "Couldn't delete database %s. Here's why: %s: %s",
                name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise


    def delete_crawler(self, name):
        """
        Deletes a crawler.

        :param name: The name of the crawler to delete.
        """
        try:
            self.glue_client.delete_crawler(Name=name)
        except ClientError as err:
            logger.error(
                "Couldn't delete crawler %s. Here's why: %s: %s",
                name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise
```

Create a class that runs the scenario.

```
class GlueCrawlerJobScenario:
    """
    Encapsulates a scenario that shows how to create an AWS Glue crawler and job
 and use
    them to transform data from CSV to JSON format.
    """

    def __init__(self, glue_client, glue_service_role, glue_bucket):
        """
        :param glue_client: A Boto3 AWS Glue client.
        :param glue_service_role: An AWS Identity and Access Management (IAM)
 role
                                  that AWS Glue can assume to gain access to the
                                  resources it requires.
        :param glue_bucket: An S3 bucket that can hold a job script and output
 data
                            from AWS Glue job runs.
        """
        self.glue_client = glue_client
        self.glue_service_role = glue_service_role
        self.glue_bucket = glue_bucket

    @staticmethod
    def wait(seconds, tick=12):
        """
        Waits for a specified number of seconds, while also displaying an
 animated
        spinner.

        :param seconds: The number of seconds to wait.
        :param tick: The number of frames per second used to animate the spinner.
        """
        progress = "|/-\\"
        waited = 0
        while waited < seconds:
            for frame in range(tick):
                sys.stdout.write(f"\r{progress[frame % len(progress)]}")
                sys.stdout.flush()
                time.sleep(1 / tick)
            waited += 1

    def upload_job_script(self, job_script):
        """
        Uploads a Python ETL script to an S3 bucket. The script is used by the 
AWS Glue
        job to transform data.

        :param job_script: The relative path to the job script.
        """
        try:
            self.glue_bucket.upload_file(Filename=job_script, Key=job_script)
            print(f"Uploaded job script '{job_script}' to the example bucket.")
        except S3UploadFailedError as err:
            logger.error("Couldn't upload job script. Here's why: %s", err)
            raise

    def run(self, crawler_name, db_name, db_prefix, data_source, job_script, 
job_name):
        """
        Runs the scenario. This is an interactive experience that runs at a 
command
        prompt and asks you for input throughout.

        :param crawler_name: The name of the crawler used in the scenario. If the
                             crawler does not exist, it is created.
        :param db_name: The name to give the metadata database created by the 
crawler.
        :param db_prefix: The prefix to give tables added to the database by the
                          crawler.
        :param data_source: The location of the data source that is targeted by 
the
                            crawler and extracted during job runs.
        :param job_script: The job script that is used to transform data during 
job
                           runs.
        :param job_name: The name to give the job definition that is created 
during the
                         scenario.
        """
        wrapper = GlueWrapper(self.glue_client)
        print(f"Checking for crawler {crawler_name}.")
        crawler = wrapper.get_crawler(crawler_name)
        if crawler is None:
            print(f"Creating crawler {crawler_name}.")
            wrapper.create_crawler(
                crawler_name,
                self.glue_service_role.arn,
                db_name,
                db_prefix,
                data_source,
            )
            print(f"Created crawler {crawler_name}.")
            crawler = wrapper.get_crawler(crawler_name)
        pprint(crawler)
        print("-" * 88)

        print(
            f"When you run the crawler, it crawls data stored in {data_source}
 and "
            f"creates a metadata database in the AWS Glue Data Catalog that
 describes "
            f"the data in the data source."
        )
        print("In this example, the source data is in CSV format.")
        ready = False
        while not ready:
            ready = Question.ask_question(
                "Ready to start the crawler? (y/n) ", Question.is_yesno
            )
        wrapper.start_crawler(crawler_name)
        print("Let's wait for the crawler to run. This typically takes a few
 minutes.")
        crawler_state = None
        while crawler_state != "READY":
            self.wait(10)
            crawler = wrapper.get_crawler(crawler_name)
            crawler_state = crawler["State"]
            print(f"Crawler is {crawler['State']}.")
        print("-" * 88)

        database = wrapper.get_database(db_name)
        print(f"The crawler created database {db_name}:")
        pprint(database)
        print(f"The database contains these tables:")
        tables = wrapper.get_tables(db_name)
        for index, table in enumerate(tables):
            print(f"\t{index + 1}. {table['Name']}")
        table_index = Question.ask_question(
            f"Enter the number of a table to see more detail: ",
            Question.is_int,
            Question.in_range(1, len(tables)),
        )
        pprint(tables[table_index - 1])
        print("-" * 88)

        print(f"Creating job definition {job_name}.")
        wrapper.create_job(
            job_name,
            "Getting started example job.",
            self.glue_service_role.arn,
            f"s3://{self.glue_bucket.name}/{job_script}",
        )
        print("Created job definition.")
        print(
            f"When you run the job, it extracts data from {data_source},
 transforms it "
            f"by using the {job_script} script, and loads the output into "
            f"S3 bucket {self.glue_bucket.name}."
        )
        print(
            "In this example, the data is transformed from CSV to JSON, and only
 a few "
            "fields are included in the output."
        )
        job_run_status = None
        if Question.ask_question(f"Ready to run? (y/n) ", Question.is_yesno):
            job_run_id = wrapper.start_job_run(
                job_name, db_name, tables[0]["Name"], self.glue_bucket.name
            )
            print(f"Job {job_name} started. Let's wait for it to run.")
            while job_run_status not in ["SUCCEEDED", "STOPPED", "FAILED",
 "TIMEOUT"]:
                self.wait(10)
                job_run = wrapper.get_job_run(job_name, job_run_id)
                job_run_status = job_run["JobRunState"]
                print(f"Job {job_name}/{job_run_id} is {job_run_status}.")
        print("-" * 88)

        if job_run_status == "SUCCEEDED":
            print(
                f"Data from your job run is stored in your S3 bucket
 '{self.glue_bucket.name}':"
            )
            try:
                keys = [
                    obj.key for obj in
 self.glue_bucket.objects.filter(Prefix="run-")
                ]
                for index, key in enumerate(keys):
                    print(f"\t{index + 1}: {key}")
                lines = 4
                key_index = Question.ask_question(
                    f"Enter the number of a block to download it and see the
 first {lines} "
                    f"lines of JSON output in the block: ",
                    Question.is_int,
                    Question.in_range(1, len(keys)),
                )
                job_data = io.BytesIO()
                self.glue_bucket.download_fileobj(keys[key_index - 1], job_data)
                job_data.seek(0)
                for _ in range(lines):
                    print(job_data.readline().decode("utf-8"))
            except ClientError as err:
                logger.error(
                    "Couldn't get job run data. Here's why: %s: %s",
                    err.response["Error"]["Code"],
                    err.response["Error"]["Message"],
                )
                raise
            print("-" * 88)

        job_names = wrapper.list_jobs()
        if job_names:
            print(f"Your account has {len(job_names)} jobs defined:")
            for index, job_name in enumerate(job_names):
                print(f"\t{index + 1}. {job_name}")
            job_index = Question.ask_question(
                f"Enter a number between 1 and {len(job_names)} to see the list
 of runs for "
```

```
                f"a job: ",
                Question.is_int,
                Question.in_range(1, len(job_names)),
            )
            job_runs = wrapper.get_job_runs(job_names[job_index - 1])
            if job_runs:
                print(f"Found {len(job_runs)} runs for job {job_names[job_index -
 1]}:")
                for index, job_run in enumerate(job_runs):
                    print(
                        f"\t{index + 1}. {job_run['JobRunState']} on "
                        f"{job_run['CompletedOn']:%Y-%m-%d %H:%M:%S}"
                    )
                run_index = Question.ask_question(
                    f"Enter a number between 1 and {len(job_runs)} to see details
 for a run: ",
                    Question.is_int,
                    Question.in_range(1, len(job_runs)),
                )
                pprint(job_runs[run_index - 1])
            else:
                print(f"No runs found for job {job_names[job_index - 1]}")
        else:
            print("Your account doesn't have any jobs defined.")
        print("-" * 88)

        print(
            f"Let's clean up. During this example we created job definition
 '{job_name}'."
        )
        if Question.ask_question(
            "Do you want to delete the definition and all runs? (y/n) ",
            Question.is_yesno,
        ):
            wrapper.delete_job(job_name)
            print(f"Job definition '{job_name}' deleted.")
        tables = wrapper.get_tables(db_name)
        print(f"We also created database '{db_name}' that contains these
 tables:")
        for table in tables:
            print(f"\t{table['Name']}")
        if Question.ask_question(
            "Do you want to delete the tables and the database? (y/n) ",
            Question.is_yesno,
        ):
            for table in tables:
                wrapper.delete_table(db_name, table["Name"])
                print(f"Deleted table {table['Name']}.")
            wrapper.delete_database(db_name)
            print(f"Deleted database {db_name}.")
        print(f"We also created crawler '{crawler_name}'.")
        if Question.ask_question(
            "Do you want to delete the crawler? (y/n) ", Question.is_yesno
        ):
            wrapper.delete_crawler(crawler_name)
            print(f"Deleted crawler {crawler_name}.")
        print("-" * 88)


def parse_args(args):
    """
    Parse command line arguments.

    :param args: The command line arguments.
    :return: The parsed arguments.
    """
    parser = argparse.ArgumentParser(
        description="Runs the AWS Glue getting started with crawlers and jobs
 scenario. "
        "Before you run this scenario, set up scaffold resources by running "
        "'python scaffold.py deploy'."
    )
    parser.add_argument(
        "role_name",
        help="The name of an IAM role that AWS Glue can assume. This role must
 grant access "
        "to Amazon S3 and to the permissions granted by the AWSGlueServiceRole "
        "managed policy.",
    )
    parser.add_argument(
        "bucket_name",
        help="The name of an S3 bucket that AWS Glue can access to get the job
 script and "
        "put job results.",
    )
    parser.add_argument(
        "--job_script",
        default="flight_etl_job_script.py",
        help="The name of the job script file that is used in the scenario.",
    )
    return parser.parse_args(args)


def main():
    args = parse_args(sys.argv[1:])
    try:
        print("-" * 88)
        print(
            "Welcome to the AWS Glue getting started with crawlers and jobs
 scenario."
        )
        print("-" * 88)
        scenario = GlueCrawlerJobScenario(
            boto3.client("glue"),
            boto3.resource("iam").Role(args.role_name),
            boto3.resource("s3").Bucket(args.bucket_name),
        )
        scenario.upload_job_script(args.job_script)
        scenario.run(
            "doc-example-crawler",
            "doc-example-database",
            "doc-example-",
            "s3://crawler-public-us-east-1/flight/2016/csv",
            args.job_script,
            "doc-example-job",
        )
        print("-" * 88)
        print(
            "To destroy scaffold resources, including the IAM role and S3 bucket
 "
            "used in this scenario, run 'python scaffold.py destroy'."
        )
        print("\nThanks for watching!")
        print("-" * 88)
    except Exception:
        logging.exception("Something went wrong with the example.")
```

Create an ETL script that is used by AWS Glue to extract, transform, and load data during job runs.

```
import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job

"""
These custom arguments must be passed as Arguments to the StartJobRun request.
    --input_database    The name of a metadata database that is contained in
 your
                        AWS Glue Data Catalog and that contains tables that
 describe
                        the data to be processed.
    --input_table       The name of a table in the database that describes the
 data to
                        be processed.
    --output_bucket_url An S3 bucket that receives the transformed output data.
"""
args = getResolvedOptions(
    sys.argv, ["JOB_NAME", "input_database", "input_table", "output_bucket_url"]
)
sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
job.init(args["JOB_NAME"], args)

# Script generated for node S3 Flight Data.
S3FlightData_node1 = glueContext.create_dynamic_frame.from_catalog(
    database=args["input_database"],
    table_name=args["input_table"],
    transformation_ctx="S3FlightData_node1",
)

# This mapping performs two main functions:
# 1. It simplifies the output by removing most of the fields from the data.
# 2. It renames some fields. For example, `fl_date` is renamed to `flight_date`.
ApplyMapping_node2 = ApplyMapping.apply(
    frame=S3FlightData_node1,
    mappings=[
        ("year", "long", "year", "long"),
        ("month", "long", "month", "tinyint"),
        ("day_of_month", "long", "day", "tinyint"),
        ("fl_date", "string", "flight_date", "string"),
        ("carrier", "string", "carrier", "string"),
        ("fl_num", "long", "flight_num", "long"),
        ("origin_city_name", "string", "origin_city_name", "string"),
        ("origin_state_abr", "string", "origin_state_abr", "string"),
        ("dest_city_name", "string", "dest_city_name", "string"),
        ("dest_state_abr", "string", "dest_state_abr", "string"),
        ("dep_time", "long", "departure_time", "long"),
        ("wheels_off", "long", "wheels_off", "long"),
        ("wheels_on", "long", "wheels_on", "long"),
        ("arr_time", "long", "arrival_time", "long"),
        ("mon", "string", "mon", "string"),
    ],
    transformation_ctx="ApplyMapping_node2",
)

# Script generated for node Revised Flight Data.
RevisedFlightData_node3 = glueContext.write_dynamic_frame.from_options(
    frame=ApplyMapping_node2,
    connection_type="s3",
    format="json",
    connection_options={"path": args["output_bucket_url"], "partitionKeys": []},
    transformation_ctx="RevisedFlightData_node3",
)

job.commit()
```

- For API details, see the following topics in *AWS SDK for Python (Boto3) API Reference*.
    - CreateCrawler
    - CreateJob
    - DeleteCrawler
    - DeleteDatabase
    - DeleteJob
    - DeleteTable
    - GetCrawler

- GetDatabase
- GetDatabases
- GetJob
- GetJobRun
- GetJobRuns
- GetTables
- ListJobs
- StartCrawler
- StartJobRun

Ruby

SDK for Ruby

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

Create a class that wraps AWS Glue functions used in the scenario.

```
# The `GlueWrapper` class serves as a wrapper around the AWS Glue API, providing
 a simplified interface for common operations.
# It encapsulates the functionality of the AWS SDK for Glue and provides methods
 for interacting with Glue crawlers, databases, tables, jobs, and S3 resources.
# The class initializes with a Glue client and a logger, allowing it to make API
 calls and log any errors or informational messages.
class GlueWrapper
  def initialize(glue_client, logger)
    @glue_client = glue_client
    @logger = logger
  end

  # Retrieves information about a specific crawler.
  #
  # @param name [String] The name of the crawler to retrieve information about.
  # @return [Aws::Glue::Types::Crawler, nil] The crawler object if found, or nil
 if not found.
  def get_crawler(name)
    @glue_client.get_crawler(name: name)
  rescue Aws::Glue::Errors::EntityNotFoundException
    @logger.info("Crawler #{name} doesn't exist.")
    false
  rescue Aws::Glue::Errors::GlueException => e
    @logger.error("Glue could not get crawler #{name}: \n#{e.message}")
    raise
  end

  # Creates a new crawler with the specified configuration.
  #
  # @param name [String] The name of the crawler.
  # @param role_arn [String] The ARN of the IAM role to be used by the crawler.
  # @param db_name [String] The name of the database where the crawler stores its
 metadata.
  # @param db_prefix [String] The prefix to be added to the names of tables that
 the crawler creates.
  # @param s3_target [String] The S3 path that the crawler will crawl.
  # @return [void]
  def create_crawler(name, role_arn, db_name, _db_prefix, s3_target)
    @glue_client.create_crawler(
      name: name,
      role: role_arn,
      database_name: db_name,
      targets: {
        s3_targets: [
          {
            path: s3_target
          }
        ]
      }
    )
  rescue Aws::Glue::Errors::GlueException => e
    @logger.error("Glue could not create crawler: \n#{e.message}")
    raise
  end

  # Starts a crawler with the specified name.
  #
  # @param name [String] The name of the crawler to start.
  # @return [void]
  def start_crawler(name)
    @glue_client.start_crawler(name: name)
  rescue Aws::Glue::Errors::ServiceError => e
    @logger.error("Glue could not start crawler #{name}: \n#{e.message}")
    raise
  end

  # Deletes a crawler with the specified name.
  #
  # @param name [String] The name of the crawler to delete.
  # @return [void]
  def delete_crawler(name)
    @glue_client.delete_crawler(name: name)
  rescue Aws::Glue::Errors::ServiceError => e
    @logger.error("Glue could not delete crawler #{name}: \n#{e.message}")
    raise
  end

  # Retrieves information about a specific database.
  #
  # @param name [String] The name of the database to retrieve information about.
  # @return [Aws::Glue::Types::Database, nil] The database object if found, or
 nil if not found.
  def get_database(name)
    response = @glue_client.get_database(name: name)
    response.database
  rescue Aws::Glue::Errors::GlueException => e
    @logger.error("Glue could not get database #{name}: \n#{e.message}")
    raise
  end

  # Retrieves a list of tables in the specified database.
  #
  # @param db_name [String] The name of the database to retrieve tables from.
  # @return [Array<Aws::Glue::Types::Table>]
  def get_tables(db_name)
    response = @glue_client.get_tables(database_name: db_name)
    response.table_list
  rescue Aws::Glue::Errors::GlueException => e
    @logger.error("Glue could not get tables #{db_name}: \n#{e.message}")
    raise
  end

  # Creates a new job with the specified configuration.
  #
  # @param name [String] The name of the job.
  # @param description [String] The description of the job.
  # @param role_arn [String] The ARN of the IAM role to be used by the job.
  # @param script_location [String] The location of the ETL script for the job.
  # @return [void]
  def create_job(name, description, role_arn, script_location)
    @glue_client.create_job(
      name: name,
      description: description,
      role: role_arn,
      command: {
        name: 'glueetl',
        script_location: script_location,
        python_version: '3'
      },
      glue_version: '3.0'
    )
  rescue Aws::Glue::Errors::GlueException => e
    @logger.error("Glue could not create job #{name}: \n#{e.message}")
    raise
  end

  # Starts a job run for the specified job.
  #
  # @param name [String] The name of the job to start the run for.
  # @param input_database [String] The name of the input database for the job.
  # @param input_table [String] The name of the input table for the job.
  # @param output_bucket_name [String] The name of the output S3 bucket for the
 job.
  # @return [String] The ID of the started job run.
  def start_job_run(name, input_database, input_table, output_bucket_name)
    response = @glue_client.start_job_run(
      job_name: name,
      arguments: {
        '--input_database': input_database,
        '--input_table': input_table,
        '--output_bucket_url': "s3://#{output_bucket_name}/"
      }
    )
    response.job_run_id
  rescue Aws::Glue::Errors::GlueException => e
    @logger.error("Glue could not start job run #{name}: \n#{e.message}")
    raise
  end

  # Retrieves a list of jobs in AWS Glue.
  #
  # @return [Aws::Glue::Types::ListJobsResponse]
  def list_jobs
    @glue_client.list_jobs
  rescue Aws::Glue::Errors::GlueException => e
    @logger.error("Glue could not list jobs: \n#{e.message}")
    raise
  end

  # Retrieves a list of job runs for the specified job.
  #
  # @param job_name [String] The name of the job to retrieve job runs for.
  # @return [Array<Aws::Glue::Types::JobRun>]
  def get_job_runs(job_name)
    response = @glue_client.get_job_runs(job_name: job_name)
    response.job_runs
  rescue Aws::Glue::Errors::GlueException => e
    @logger.error("Glue could not get job runs: \n#{e.message}")
  end

  # Retrieves data for a specific job run.
  #
  # @param job_name [String] The name of the job run to retrieve data for.
  # @return [Glue::Types::GetJobRunResponse]
  def get_job_run(job_name, run_id)
    @glue_client.get_job_run(job_name: job_name, run_id: run_id)
  rescue Aws::Glue::Errors::GlueException => e
    @logger.error("Glue could not get job runs: \n#{e.message}")
  end

  # Deletes a job with the specified name.
  #
  # @param job_name [String] The name of the job to delete.
  # @return [void]
  def delete_job(job_name)
    @glue_client.delete_job(job_name: job_name)
  rescue Aws::Glue::Errors::ServiceError => e
    @logger.error("Glue could not delete job: \n#{e.message}")
  end

  # Deletes a table with the specified name.
  #
  # @param database_name [String] The name of the catalog database in which the
 table resides.
  # @param table_name [String] The name of the table to be deleted.
  # @return [void]
  def delete_table(database_name, table_name)
    @glue_client.delete_table(database_name: database_name, name: table_name)
  rescue Aws::Glue::Errors::ServiceError => e
    @logger.error("Glue could not delete job: \n#{e.message}")
  end

  # Removes a specified database from a Data Catalog.
  #
  # @param database_name [String] The name of the database to delete.
  # @return [void]
  def delete_database(database_name)
    @glue_client.delete_database(name: database_name)
  rescue Aws::Glue::Errors::ServiceError => e
    @logger.error("Glue could not delete database: \n#{e.message}")
  end

  # Uploads a job script file to an S3 bucket.
  #
  # @param file_path [String] The local path of the job script file.
  # @param bucket_resource [Aws::S3::Bucket] The S3 bucket resource to upload the
 file to.
  # @return [void]
  def upload_job_script(file_path, bucket_resource)
    File.open(file_path) do |file|
      bucket_resource.client.put_object({
                                          body: file,
                                          bucket: bucket_resource.name,
                                          key: file_path
                                        })
    end
  rescue Aws::S3::Errors::S3UploadFailedError => e
    @logger.error("S3 could not upload job script: \n#{e.message}")
    raise
  end
end
```

Create a class that runs the scenario.

```
class GlueCrawlerJobScenario
  def initialize(glue_client, glue_service_role, glue_bucket, logger)
    @glue_client = glue_client
    @glue_service_role = glue_service_role
    @glue_bucket = glue_bucket
    @logger = logger
  end

  def run(crawler_name, db_name, db_prefix, data_source, job_script, job_name)
    wrapper = GlueWrapper.new(@glue_client, @logger)
    setup_crawler(wrapper, crawler_name, db_name, db_prefix, data_source)
    query_database(wrapper, crawler_name, db_name)
    create_and_run_job(wrapper, job_script, job_name, db_name)
  end

  private

  def setup_crawler(wrapper, crawler_name, db_name, db_prefix, data_source)
    new_step(1, 'Create a crawler')
    crawler = wrapper.get_crawler(crawler_name)
    unless crawler
      puts "Creating crawler #{crawler_name}."
      wrapper.create_crawler(crawler_name, @glue_service_role.arn, db_name,
 db_prefix, data_source)
      puts "Successfully created #{crawler_name}."
    end
    wrapper.start_crawler(crawler_name)
    monitor_crawler(wrapper, crawler_name)
  end

  def monitor_crawler(wrapper, crawler_name)
    new_step(2, 'Monitor Crawler')
    crawler_state = nil
    until crawler_state == 'READY'
      custom_wait(15)
      crawler = wrapper.get_crawler(crawler_name)
      crawler_state = crawler[0]['state']
      print "Crawler status: #{crawler_state}".yellow
    end
  end

  def query_database(wrapper, _crawler_name, db_name)
    new_step(3, 'Query the database.')
    wrapper.get_database(db_name)
    puts "The crawler created database #{db_name}:"
    puts "Database contains tables: #{wrapper.get_tables(db_name).map { |t|
 t['name'] }}"
  end

  def create_and_run_job(wrapper, job_script, job_name, db_name)
    new_step(4, 'Create and run job.')
    wrapper.upload_job_script(job_script, @glue_bucket)
    wrapper.create_job(job_name, 'ETL Job', @glue_service_role.arn, "s3://
#{@glue_bucket.name}/#{job_script}")
    run_job(wrapper, job_name, db_name)
  end

  def run_job(wrapper, job_name, db_name)
    new_step(5, 'Run the job.')
    wrapper.start_job_run(job_name, db_name, wrapper.get_tables(db_name)[0]
['name'], @glue_bucket.name)
    job_run_status = nil
    until %w[SUCCEEDED FAILED STOPPED].include?(job_run_status)
      custom_wait(10)
      job_run = wrapper.get_job_runs(job_name)
      job_run_status = job_run[0]['job_run_state']
      print "Job #{job_name} status: #{job_run_status}".yellow
    end
  end
end

def main
  banner('../../helpers/banner.txt')
  puts 'Starting AWS Glue demo...'

  # Load resource names from YAML.
  resource_names = YAML.load_file('resource_names.yaml')

  # Setup services and resources.
  iam_role = Aws::IAM::Resource.new(region: 'us-
east-1').role(resource_names['glue_service_role'])
  s3_bucket = Aws::S3::Resource.new(region: 'us-
east-1').bucket(resource_names['glue_bucket'])

  # Instantiate scenario and run.
  scenario = GlueCrawlerJobScenario.new(Aws::Glue::Client.new(region: 'us-
east-1'), iam_role, s3_bucket, @logger)
```

```
  random_suffix = rand(10**4)
  scenario.run("crawler-#{random_suffix}", "db-#{random_suffix}", "prefix-
#{random_suffix}-", 's3://data_source',
               'job_script.py', "job-#{random_suffix}")

  puts 'Demo complete.'
end
```

Create an ETL script that is used by AWS Glue to extract, transform, and load data during job runs.

```
import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job

"""
These custom arguments must be passed as Arguments to the StartJobRun request.
    --input_database    The name of a metadata database that is contained in
 your
                        AWS Glue Data Catalog and that contains tables that
 describe
                        the data to be processed.
    --input_table       The name of a table in the database that describes the
 data to
                        be processed.
    --output_bucket_url An S3 bucket that receives the transformed output data.
"""
args = getResolvedOptions(
    sys.argv, ["JOB_NAME", "input_database", "input_table", "output_bucket_url"]
)
sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
job.init(args["JOB_NAME"], args)

# Script generated for node S3 Flight Data.
S3FlightData_node1 = glueContext.create_dynamic_frame.from_catalog(
    database=args["input_database"],
    table_name=args["input_table"],
    transformation_ctx="S3FlightData_node1",
)

# This mapping performs two main functions:
# 1. It simplifies the output by removing most of the fields from the data.
# 2. It renames some fields. For example, `fl_date` is renamed to `flight_date`.
ApplyMapping_node2 = ApplyMapping.apply(
    frame=S3FlightData_node1,
    mappings=[
        ("year", "long", "year", "long"),
        ("month", "long", "month", "tinyint"),
        ("day_of_month", "long", "day", "tinyint"),
        ("fl_date", "string", "flight_date", "string"),
        ("carrier", "string", "carrier", "string"),
        ("fl_num", "long", "flight_num", "long"),
        ("origin_city_name", "string", "origin_city_name", "string"),
        ("origin_state_abr", "string", "origin_state_abr", "string"),
        ("dest_city_name", "string", "dest_city_name", "string"),
        ("dest_state_abr", "string", "dest_state_abr", "string"),
        ("dep_time", "long", "departure_time", "long"),
        ("wheels_off", "long", "wheels_off", "long"),
        ("wheels_on", "long", "wheels_on", "long"),
        ("arr_time", "long", "arrival_time", "long"),
        ("mon", "string", "mon", "string"),
    ],
    transformation_ctx="ApplyMapping_node2",
)

# Script generated for node Revised Flight Data.
RevisedFlightData_node3 = glueContext.write_dynamic_frame.from_options(
    frame=ApplyMapping_node2,
    connection_type="s3",
    format="json",
    connection_options={"path": args["output_bucket_url"], "partitionKeys": []},
    transformation_ctx="RevisedFlightData_node3",
)

job.commit()
```

- For API details, see the following topics in *AWS SDK for Ruby API Reference*.
  - CreateCrawler

- CreateJob
- DeleteCrawler
- DeleteDatabase
- DeleteJob
- DeleteTable
- GetCrawler
- GetDatabase
- GetDatabases
- GetJob
- GetJobRun
- GetJobRuns
- GetTables
- ListJobs
- StartCrawler
- StartJobRun

Rust

SDK for Rust

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

Create and run a crawler that crawls a public Amazon Simple Storage Service (Amazon S3) bucket and generates a metadata database that describes the CSV-formatted data it finds.

```
        let create_crawler = glue
            .create_crawler()
            .name(self.crawler())
            .database_name(self.database())
            .role(self.iam_role.expose_secret())
            .targets(
                CrawlerTargets::builder()
                    .s3_targets(S3Target::builder().path(CRAWLER_TARGET).build())
                    .build(),
            )
            .send()
            .await;

        match create_crawler {
            Err(err) => {
                let glue_err: aws_sdk_glue::Error = err.into();
                match glue_err {
                    aws_sdk_glue::Error::AlreadyExistsException(_) => {
                        info!("Using existing crawler");
                        Ok(())
                    }
                    _ => Err(GlueMvpError::GlueSdk(glue_err)),
                }
            }
            Ok(_) => Ok(()),
        }?;

        let start_crawler =
 glue.start_crawler().name(self.crawler()).send().await;

        match start_crawler {
            Ok(_) => Ok(()),
            Err(err) => {
                let glue_err: aws_sdk_glue::Error = err.into();
                match glue_err {
                    aws_sdk_glue::Error::CrawlerRunningException(_) => Ok(()),
                    _ => Err(GlueMvpError::GlueSdk(glue_err)),
                }
            }
        }?;
```

List information about databases and tables in your AWS Glue Data Catalog.

```
        let database = glue
            .get_database()
            .name(self.database())
            .send()
            .await
            .map_err(GlueMvpError::from_glue_sdk)?
            .to_owned();
        let database = database
            .database()
            .ok_or_else(|| GlueMvpError::Unknown("Could not find 
database".into()))?;

        let tables = glue
            .get_tables()
            .database_name(self.database())
            .send()
            .await
            .map_err(GlueMvpError::from_glue_sdk)?;

        let tables = tables.table_list();
```

Create and run a job that extracts CSV data from the source Amazon S3 bucket, transforms it by removing and renaming fields, and loads JSON-formatted output into another Amazon S3 bucket.

```
        let create_job = glue
            .create_job()
            .name(self.job())
            .role(self.iam_role.expose_secret())
            .command(
                JobCommand::builder()
                    .name("glueetl")
                    .python_version("3")
                    .script_location(format!("s3://{}/job.py", self.bucket()))
                    .build(),
            )
            .glue_version("3.0")
            .send()
            .await
            .map_err(GlueMvpError::from_glue_sdk)?;

        let job_name = create_job.name().ok_or_else(|| {
            GlueMvpError::Unknown("Did not get job name after creating 
job".into())
        })?;

        let job_run_output = glue
```

```
            .start_job_run()
            .job_name(self.job())
            .arguments("--input_database", self.database())
            .arguments(
                "--input_table",
                self.tables
                    .first()
                    .ok_or_else(|| GlueMvpError::Unknown("Missing crawler 
table".into()))?
                    .name(),
            )
            .arguments("--output_bucket_url", self.bucket())
            .send()
            .await
            .map_err(GlueMvpError::from_glue_sdk)?;

        let job = job_run_output
            .job_run_id()
            .ok_or_else(|| GlueMvpError::Unknown("Missing run id from just 
started job".into()))?
            .to_string();
```

Delete all resources created by the demo.

```
        glue.delete_job()
            .job_name(self.job())
            .send()
            .await
            .map_err(GlueMvpError::from_glue_sdk)?;

        for t in &self.tables {
            glue.delete_table()
                .name(t.name())
                .database_name(self.database())
                .send()
                .await
                .map_err(GlueMvpError::from_glue_sdk)?;
        }

        glue.delete_database()
            .name(self.database())
            .send()
            .await
            .map_err(GlueMvpError::from_glue_sdk)?;

        glue.delete_crawler()
            .name(self.crawler())
            .send()
            .await
            .map_err(GlueMvpError::from_glue_sdk)?;
```

- For API details, see the following topics in *AWS SDK for Rust API reference*.
  - CreateCrawler
  - CreateJob
  - DeleteCrawler
  - DeleteDatabase
  - DeleteJob
  - DeleteTable
  - GetCrawler
  - GetDatabase
  - GetDatabases
  - GetJob
  - GetJobRun
  - GetJobRuns
  - GetTables
  - ListJobs
  - StartCrawler
  - StartJobRun

Swift

SDK for Swift

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

The `Package.swift` file.

```
// swift-tools-version: 5.9
//
// The swift-tools-version declares the minimum version of Swift required to
// build this package.

import PackageDescription

let package = Package(
    name: "glue-scenario",
    // Let Xcode know the minimum Apple platforms supported.
    platforms: [
        .macOS(.v13),
        .iOS(.v15)
    ],
    dependencies: [
        // Dependencies declare other packages that this package depends on.
        .package(
            url: "https://github.com/awslabs/aws-sdk-swift",
            from: "1.0.0"),
        .package(
            url: "https://github.com/apple/swift-argument-parser.git",
            branch: "main"
        )
    ],
    targets: [
        // Targets are the basic building blocks of a package, defining a module
 or a test suite.
        // Targets can depend on other targets in this package and products
        // from dependencies.
        .executableTarget(
            name: "glue-scenario",
            dependencies: [
                .product(name: "AWSGlue", package: "aws-sdk-swift"),
                .product(name: "AWSS3", package: "aws-sdk-swift"),
                .product(name: "ArgumentParser", package: "swift-argument-
parser")
            ],
            path: "Sources")

    ]
)
```

The Swift code file, `entry.swift`.

```
// An example that shows how to use the AWS SDK for Swift to demonstrate
// creating and using crawlers and jobs using AWS Glue.
//
// 0. Upload the Python job script to Amazon S3 so it can be used when
//    calling `startJobRun()` later.
// 1. Create a crawler, pass it the IAM role and the URL of the public Amazon
//    S3 bucket that contains the source data:
//    s3://crawler-public-us-east-1/flight/2016/csv.
// 2. Start the crawler. This takes time, so after starting it, use a loop
//    that calls `getCrawler()` until the state is "READY".
// 3. Get the database created by the crawler, and the tables in the
//    database. Display them to the user.
// 4. Create a job. Pass it the IAM role and the URL to a Python ETL script
//    previously uploaded to the user's S3 bucket.
// 5. Start a job run, passing the following custom arguments. These are
//    expected by the ETL script, so must exactly match.
//    * `--input_database: <name of the database created by the crawler>`
//    * `--input_table: <name of the table created by the crawler>`
//    * `--output_bucket_url: <URL to the scaffold bucket created for the
//      user>`
// 6. Loop and get the job run until it returns one of the following states:
//    "SUCCEEDED", "STOPPED", "FAILED", or "TIMEOUT".
// 7. Output data is stored in a group of files in the user's S3 bucket.
//    Either direct the user to their location or download a file and display
//    the results inline.
// 8. List the jobs for the user's account.
// 9. Get job run details for a job run.
// 10. Delete the demo job.
// 11. Delete the database and tables created by the example.
// 12. Delete the crawler created by the example.

import ArgumentParser
import AWSS3
import Foundation
import Smithy

import AWSClientRuntime
import AWSGlue

struct ExampleCommand: ParsableCommand {
    @Option(help: "The AWS IAM role to use for AWS Glue calls.")
    var role: String

    @Option(help: "The Amazon S3 bucket to use for this example.")
    var bucket: String

    @Option(help: "The Amazon S3 URL of the data to crawl.")
    var s3url: String = "s3://crawler-public-us-east-1/flight/2016/csv"

    @Option(help: "The Python script to run as a job with AWS Glue.")
    var script: String = "./flight_etl_job_script.py"

    @Option(help: "The AWS Region to run AWS API calls in.")
    var awsRegion = "us-east-1"

    @Option(help: "A prefix string to use when naming tables.")
    var tablePrefix = "swift-glue-basics-table"

    @Option(
        help: ArgumentHelp("The level of logging for the Swift SDK to perform."),
        completion: .list([
            "critical",
            "debug",
            "error",
            "info",
            "notice",
            "trace",
            "warning"
        ])
    )
    var logLevel: String = "error"

    static var configuration = CommandConfiguration(
        commandName: "glue-scenario",
        abstract: """
        Demonstrates various features of AWS Glue.
        """,
        discussion: """
        An example showing how to use AWS Glue to create, run, and monitor
        crawlers and jobs.
        """
    )

    /// Generate and return a unique file name that begins with the specified
    /// string.
    ///
    /// - Parameters:
    ///   - prefix: Text to use at the beginning of the returned name.
    ///
    /// - Returns: A string containing a unique filename that begins with the
    ///   specified `prefix`.
    ///
    /// The returned name uses a random number between 1 million and 1 billion to
    /// provide reasonable certainty of uniqueness for the purposes of this
    /// example.
    func tempName(prefix: String) -> String {
        return "\(prefix)-\(Int.random(in: 1000000..<1000000000))"
    }

    /// Upload a file to an Amazon S3 bucket.
    ///
    /// - Parameters:
    ///   - s3Client: The S3 client to use when uploading the file.
    ///   - path: The local path of the source file to upload.
    ///   - toBucket: The name of the S3 bucket into which to upload the file.
    ///   - key: The key (name) to give the file in the S3 bucket.
    ///
    /// - Returns: `true` if the file is uploaded successfully, otherwise
 `false`.
    func uploadFile(s3Client: S3Client, path: String, toBucket: String, key:
 String) async -> Bool {
        do {
            let fileData: Data = try Data(contentsOf: URL(fileURLWithPath: path))
            let dataStream = ByteStream.data(fileData)
            _ = try await s3Client.putObject(
                input: PutObjectInput(
                    body: dataStream,
                    bucket: toBucket,
                    key: key
                )
            )
        } catch {
            print("*** An unexpected error occurred uploading the script to the
Amazon S3 bucket \"\(bucket)\".")
            return false
        }

        return true
    }

    /// Create a new AWS Glue crawler.
    ///
    /// - Parameters:
    ///   - glueClient: An AWS Glue client to use for the crawler.
    ///   - crawlerName: A name for the new crawler.
    ///   - iamRole: The name of an Amazon IAM role for the crawler to use.
    ///   - s3Path: The path of an Amazon S3 folder to use as a target location.
    ///   - cronSchedule: A `cron` schedule indicating when to run the crawler.
    ///   - databaseName: The name of an AWS Glue database to operate on.
    ///
    /// - Returns: `true` if the crawler is created successfully, otherwise
`false`.
    func createCrawler(glueClient: GlueClient, crawlerName: String, iamRole:
String,
                       s3Path: String, cronSchedule: String, databaseName:
String) async -> Bool {
        let s3Target = GlueClientTypes.S3Target(path: s3url)
        let targetList = GlueClientTypes.CrawlerTargets(s3Targets: [s3Target])

        do {
            _ = try await glueClient.createCrawler(
                input: CreateCrawlerInput(
                    databaseName: databaseName,
                    description: "Created by the AWS SDK for Swift Scenario
Example for AWS Glue.",
                    name: crawlerName,
                    role: iamRole,
                    schedule: cronSchedule,
                    tablePrefix: tablePrefix,
                    targets: targetList
                )
            )
        } catch _ as AlreadyExistsException {
            print("*** A crawler named \"\(crawlerName)\" already exists.")
            return false
        } catch _ as OperationTimeoutException {
            print("*** The attempt to create the AWS Glue crawler timed out.")
            return false
        } catch {
            print("*** An unexpected error occurred creating the AWS Glue 
 crawler: \(error.localizedDescription)")
            return false
        }

        return true
    }

    /// Delete an AWS Glue crawler.
    ///
    /// - Parameters:
    ///   - glueClient: The AWS Glue client to use.
    ///   - name: The name of the crawler to delete.
    ///
    /// - Returns: `true` if successful, otherwise `false`.
    func deleteCrawler(glueClient: GlueClient, name: String) async -> Bool {
        do {
            _ = try await glueClient.deleteCrawler(
                input: DeleteCrawlerInput(name: name)
            )
        } catch {
            return false
        }
        return true
    }

    /// Start running an AWS Glue crawler.
    ///
    /// - Parameters:
    ///   - glueClient: The AWS Glue client to use when starting the crawler.
    ///   - name: The name of the crawler to start running.
    ///
    /// - Returns: `true` if the crawler is started successfully, otherwise 
 `false`.
    func startCrawler(glueClient: GlueClient, name: String) async -> Bool {
        do {
            _ = try await glueClient.startCrawler(
                input: StartCrawlerInput(name: name)
            )
        } catch {
            print("*** An unexpected error occurred starting the crawler.")
            return false
        }

        return true
    }

    /// Get the state of the specified AWS Glue crawler.
    ///
    /// - Parameters:
    ///   - glueClient: The AWS Glue client to use.
    ///   - name: The name of the crawler whose state should be returned.
    ///
    /// - Returns: A `GlueClientTypes.CrawlerState` value describing the
    ///   state of the crawler.
    func getCrawlerState(glueClient: GlueClient, name: String) async ->
 GlueClientTypes.CrawlerState {
        do {
            let output = try await glueClient.getCrawler(
                input: GetCrawlerInput(name: name)
            )

            // If the crawler or its state is `nil`, report that the crawler
            // is stopping. This may not be what you want for your
            // application but it works for this one!

            guard let crawler = output.crawler else {
                return GlueClientTypes.CrawlerState.stopping
            }
            guard let state = crawler.state else {
                return GlueClientTypes.CrawlerState.stopping
            }
            return state
        } catch {
            return GlueClientTypes.CrawlerState.stopping
        }
    }

    /// Wait until the specified crawler is ready to run.
    ///
    /// - Parameters:
    ///   - glueClient: The AWS Glue client to use.
    ///   - name: The name of the crawler to wait for.
    ///
    /// - Returns: `true` if the crawler is ready, `false` if the client is
    ///   stopping (and will therefore never be ready).
    func waitUntilCrawlerReady(glueClient: GlueClient, name: String) async ->
 Bool {
        while true {
            let state = await getCrawlerState(glueClient: glueClient, name: name)

            if state == .ready {
                return true
            } else if state == .stopping {
                return false
            }

            // Wait four seconds before trying again.

            do {
                try await Task.sleep(for: .seconds(4))
            } catch {
                print("*** Error pausing the task.")
            }
        }
    }

    /// Create a new AWS Glue job.
    ///
    /// - Parameters:
    ///   - glueClient: The AWS Glue client to use.
    ///   - jobName: The name to give the new job.
    ///   - role: The IAM role for the job to use when accessing AWS services.
    ///   - scriptLocation: The AWS S3 URI of the script to be run by the job.
    ///
    /// - Returns: `true` if the job is created successfully, otherwise `false`.
    func createJob(glueClient: GlueClient, name jobName: String, role: String,
                   scriptLocation: String) async -> Bool {
        let command = GlueClientTypes.JobCommand(
            name: "glueetl",
            pythonVersion: "3",
            scriptLocation: scriptLocation
        )
        do {
            _ = try await glueClient.createJob(
                input: CreateJobInput(
                    command: command,
                    description: "Created by the AWS SDK for Swift Glue basic
 scenario example.",
                    glueVersion: "3.0",
                    name: jobName,
                    numberOfWorkers: 10,
                    role: role,
                    workerType: .g1x
                )
            )
        } catch {
            return false
        }
        return true
    }

    /// Return a list of the AWS Glue jobs listed on the user's account.
    ///
    /// - Parameters:
    ///   - glueClient: The AWS Glue client to use.
    ///   - maxJobs: The maximum number of jobs to return (default: 100).
    ///
    /// - Returns: An array of strings listing the names of all available AWS
    ///   Glue jobs.
    func listJobs(glueClient: GlueClient, maxJobs: Int = 100) async -> [String] {
        var jobList: [String] = []
        var nextToken: String?

        repeat {
            do {
                let output = try await glueClient.listJobs(
                    input: ListJobsInput(
                        maxResults: maxJobs,
                        nextToken: nextToken
                    )
                )

                guard let jobs = output.jobNames else {
                    return jobList
                }
                jobList = jobList + jobs
                nextToken = output.nextToken
            } catch {
                return jobList
            }
        } while (nextToken != nil)

        return jobList
    }

    /// Delete an AWS Glue job.
    ///
    /// - Parameters:
    ///   - glueClient: The AWS Glue client to use.
    ///   - jobName: The name of the job to delete.
    ///
    /// - Returns: `true` if the job is successfully deleted, otherwise `false`.
    func deleteJob(glueClient: GlueClient, name jobName: String) async -> Bool {
        do {
            _ = try await glueClient.deleteJob(
                input: DeleteJobInput(jobName: jobName)
            )
        } catch {
            return false
        }
        return true
    }

    /// Create an AWS Glue database.
    ///
    /// - Parameters:
    ///   - glueClient: The AWS Glue client to use.
    ///   - databaseName: The name to give the new database.
    ///   - location: The URL of the source data to use with AWS Glue.
    ///
    /// - Returns: `true` if the database is created successfully, otherwise
 `false`.
    func createDatabase(glueClient: GlueClient, name databaseName: String,
 location: String) async -> Bool {
        let databaseInput = GlueClientTypes.DatabaseInput(
            description: "Created by the AWS SDK for Swift Glue basic scenario
 example.",
            locationUri: location,
            name: databaseName
        )

        do {
            _ = try await glueClient.createDatabase(
                input: CreateDatabaseInput(
                    databaseInput: databaseInput
                )
            )
        } catch {
            return false
        }

        return true
    }

    /// Get the AWS Glue database with the specified name.
    ///
    /// - Parameters:
    ///   - glueClient: The AWS Glue client to use.
    ///   - name: The name of the database to return.
    ///
    /// - Returns: The `GlueClientTypes.Database` object describing the
    ///   specified database, or `nil` if an error occurs or the database
    ///   isn't found.
    func getDatabase(glueClient: GlueClient, name: String) async ->
 GlueClientTypes.Database? {
        do {
            let output = try await glueClient.getDatabase(
                input: GetDatabaseInput(name: name)
            )

            return output.database
        } catch {
            return nil
        }
    }

    /// Returns a list of the tables in the specified database.
    ///
    /// - Parameters:
    ///   - glueClient: The AWS Glue client to use.
    ///   - databaseName: The name of the database whose tables are to be
    ///     returned.
    ///
    /// - Returns: An array of `GlueClientTypes.Table` objects, each
    ///   describing one table in the named database. An empty array indicates
    ///   that there are either no tables in the database, or an error
    ///   occurred before any tables could be found.
    func getTablesInDatabase(glueClient: GlueClient, databaseName: String) async
 -> [GlueClientTypes.Table] {
        var tables: [GlueClientTypes.Table] = []
        var nextToken: String?

        repeat {
            do {
                let output = try await glueClient.getTables(
                    input: GetTablesInput(
                        databaseName: databaseName,
                        nextToken: nextToken
                    )
                )

                guard let tableList = output.tableList else {
                    return tables
                }

                tables = tables + tableList
                nextToken = output.nextToken
            } catch {
                return tables
            }
        } while nextToken != nil

        return tables
    }

    /// Delete the specified database.
    ///
    /// - Parameters:
    ///   - glueClient: The AWS Glue client to use.
    ///   - databaseName: The name of the database to delete.
    ///   - deleteTables: A Bool indicating whether or not to delete the
    ///     tables in the database before attempting to delete the database.
    ///
    /// - Returns: `true` if the database (and optionally its tables) are
    ///   deleted, otherwise `false`.
    func deleteDatabase(glueClient: GlueClient, name databaseName: String,
                        withTables deleteTables: Bool = false) async -> Bool {
```

```
        if deleteTables {
            var tableNames: [String] = []

            // Get a list of the names of all of the tables in the database.

            let tableList = await self.getTablesInDatabase(glueClient: 
glueClient, databaseName: databaseName)
            for table in tableList {
                guard let name = table.name else {
                    continue
                }
                tableNames.append(name)
            }

            // Delete the tables. If there's only one table, use
            // `deleteTable()`, otherwise, use `batchDeleteTable()`. You can
            // use `batchDeleteTable()` for a single table, but this
            // demonstrates the use of `deleteTable()`.

            if tableNames.count == 1 {
                do {
                    print("    Deleting table...")
                    _ = try await glueClient.deleteTable(
                        input: DeleteTableInput(
                            databaseName: databaseName,
                            name: tableNames[0]
                        )
                    )
                } catch {
                    print("*** Unable to delete the table.")
                }
            } else {
                do {
                    print("    Deleting tables...")
                    _ = try await glueClient.batchDeleteTable(
                        input: BatchDeleteTableInput(
                            databaseName: databaseName,
                            tablesToDelete: tableNames
                        )
                    )
                } catch {
                    print("*** Unable to delete the tables.")
                }
            }
        }

        // Delete the database itself.

        do {
            print("    Deleting the database itself...")
            _ = try await glueClient.deleteDatabase(
                input: DeleteDatabaseInput(name: databaseName)
            )
        } catch {
            print("*** Unable to delete the database.")
            return false
        }
        return true
    }

    /// Start an AWS Glue job run.
    ///
    /// - Parameters:
    ///   - glueClient: The AWS Glue client to use.
    ///   - jobName: The name of the job to run.
    ///   - databaseName: The name of the AWS Glue database to run the job
 against.
    ///   - tableName: The name of the table in the database to run the job
 against.
    ///   - outputURL: The AWS S3 URI of the bucket location into which to
    ///     write the resulting output.
    ///
    /// - Returns: `true` if the job run is started successfully, otherwise
 `false`.
    func startJobRun(glueClient: GlueClient, name jobName: String, databaseName:
 String,
                     tableName: String, outputURL: String) async -> String? {
        do {
            let output = try await glueClient.startJobRun(
                input: StartJobRunInput(
                    arguments: [
                        "--input_database": databaseName,
                        "--input_table": tableName,
                        "--output_bucket_url": outputURL
                    ],
                    jobName: jobName,
                    numberOfWorkers: 10,
                    workerType: .g1x
                )
            )

            guard let id = output.jobRunId else {
                return nil
            }

            return id
        } catch {
            return nil
        }
    }

    /// Return a list of the job runs for the specified job.
    ///
    /// - Parameters:
    ///   - glueClient: The AWS Glue client to use.
    ///   - jobName: The name of the job for which to return its job runs.
    ///   - maxResults: The maximum number of job runs to return (default:
    ///     1000).
    ///
    /// - Returns: An array of `GlueClientTypes.JobRun` objects describing
    ///   each job run.
    func getJobRuns(glueClient: GlueClient, name jobName: String, maxResults:
Int? = nil) async -> [GlueClientTypes.JobRun] {
        do {
            let output = try await glueClient.getJobRuns(
                input: GetJobRunsInput(
                    jobName: jobName,
                    maxResults: maxResults
                )
            )

            guard let jobRuns = output.jobRuns else {
                print("*** No job runs found.")
                return []
            }

            return jobRuns
        } catch is EntityNotFoundException {
            print("*** The specified job name, \(jobName), doesn't exist.")
            return []
        } catch {
            print("*** Unexpected error getting job runs:")
            dump(error)
            return []
        }
    }

    /// Get information about a specific AWS Glue job run.
    ///
    /// - Parameters:
    ///   - glueClient: The AWS Glue client to use.
    ///   - jobName: The name of the job to return job run data for.
    ///   - id: The run ID of the specific job run to return.
    ///
    /// - Returns: A `GlueClientTypes.JobRun` object describing the state of
    ///   the job run, or `nil` if an error occurs.
    func getJobRun(glueClient: GlueClient, name jobName: String, id: String)
 async -> GlueClientTypes.JobRun? {
        do {
            let output = try await glueClient.getJobRun(
                input: GetJobRunInput(
                    jobName: jobName,
                    runId: id
                )
            )

            return output.jobRun
        } catch {
            return nil
        }
    }

    /// Called by ``main()`` to run the bulk of the example.
    func runAsync() async throws {
        // A name to give the Python script upon upload to the Amazon S3
        // bucket.
        let scriptName = "jobscript.py"

        // Schedule string in `cron` format, as described here:
        // https://docs.aws.amazon.com/glue/latest/dg/monitor-data-warehouse-
schedule.html
        let cron = "cron(15 12 * * ? *)"

        let glueConfig = try await GlueClient.GlueClientConfiguration(region:
 awsRegion)
        let glueClient = GlueClient(config: glueConfig)
        let s3Config = try await S3Client.S3ClientConfiguration(region:
awsRegion)
        let s3Client = S3Client(config: s3Config)

        // Create random names for things that need them.

        let crawlerName = tempName(prefix: "swift-glue-basics-crawler")
        let databaseName = tempName(prefix: "swift-glue-basics-db")

        // Create a name for the AWS Glue job.

        let jobName = tempName(prefix: "scenario-job")

        // The URL of the Python script on S3.

        let scriptURL = "s3://\(bucket)/\(scriptName)"

        print("Welcome to the AWS SDK for Swift basic scenario for AWS Glue!")

        //===================================================================
        // 0. Upload the Python script to the target bucket so it's available
        //    for use by the Amazon Glue service.
        //===================================================================

        print("Uploading the Python script: \(script) as key \(scriptName)")
        print("Destination bucket: \(bucket)")
        if !(await uploadFile(s3Client: s3Client, path: script, toBucket: bucket,
key: scriptName)) {
            return
        }

        //===================================================================
        // 1. Create the database and crawler using the randomized names
        //    generated previously.
        //===================================================================

        print("Creating database \"\(databaseName)\"...")
        if !(await createDatabase(glueClient: glueClient, name: databaseName,
location: s3url)) {
            print("*** Unable to create the database.")
            return
        }
        print("Creating crawler \"\(crawlerName)\"...")
        if !(await createCrawler(glueClient: glueClient, crawlerName: 
crawlerName,
                                iamRole: role, s3Path: s3url, cronSchedule: 
cron,
                                databaseName: databaseName)) {
            return
        }

        //=================================================================
        // 2. Start the crawler, then wait for it to be ready.
        //=================================================================

        print("Starting the crawler and waiting until it's ready...")
        if !(await startCrawler(glueClient: glueClient, name: crawlerName)) {
            _ = await deleteCrawler(glueClient: glueClient, name: crawlerName)
            return
        }

        if !(await waitUntilCrawlerReady(glueClient: glueClient, name: 
crawlerName)) {
            _ = await deleteCrawler(glueClient: glueClient, name: crawlerName)
        }

        //=================================================================
        // 3. Get the database and table created by the crawler.
        //=================================================================

        print("Getting the crawler's database...")
        let database = await getDatabase(glueClient: glueClient, name: 
databaseName)

        guard let database else {
            print("*** Unable to get the database.")
            return
        }
        print("Database URI: \(database.locationUri ?? "<unknown>")")

        let tableList = await getTablesInDatabase(glueClient: glueClient, 
databaseName: databaseName)

        print("Found \(tableList.count) table(s):")
        for table in tableList {
            print("  \(table.name ?? "<unnamed>")")
        }

        if tableList.count != 1 {
            print("*** Incorrect number of tables found. There should only be
 one.")
            _ = await deleteDatabase(glueClient: glueClient, name: databaseName,
 withTables: true)
            _ = await deleteCrawler(glueClient: glueClient, name: crawlerName)
            return
        }

        guard let tableName = tableList[0].name else {
            print("*** Table is unnamed.")
            _ = await deleteDatabase(glueClient: glueClient, name: databaseName,
 withTables: true)
            _ = await deleteCrawler(glueClient: glueClient, name: crawlerName)
            return
        }

        //=====================================================================
        // 4. Create a job.
        //=====================================================================

        print("Creating a job...")
        if !(await createJob(glueClient: glueClient, name: jobName, role: role,
                             scriptLocation: scriptURL)) {
            _ = await deleteDatabase(glueClient: glueClient, name: databaseName,
 withTables: true)
            _ = await deleteCrawler(glueClient: glueClient, name: crawlerName)
            return
        }

        //=====================================================================
        // 5. Start a job run.
        //=====================================================================

        print("Starting the job...")

        // Construct the Amazon S3 URL for the job run's output. This is in
        // the bucket specified on the command line, with a folder name that's
        // unique for this job run.

        let timeStamp = Date().timeIntervalSince1970
        let jobPath = "\(jobName)-\(Int(timeStamp))"
        let outputURL = "s3://\(bucket)/\(jobPath)"

        // Start the job run.

        let jobRunID = await startJobRun(glueClient: glueClient, name: jobName,
                                         databaseName: databaseName,
                                         tableName: tableName,
                                         outputURL: outputURL)

        guard let jobRunID else {
            print("*** Job run ID is invalid.")
            _ = await deleteJob(glueClient: glueClient, name: jobName)
            _ = await deleteDatabase(glueClient: glueClient, name: databaseName,
 withTables: true)
            _ = await deleteCrawler(glueClient: glueClient, name: crawlerName)
            return
        }

        //=====================================================================
        // 6. Wait for the job run to indicate that the run is complete.
        //=====================================================================

        print("Waiting for job run to end...")

        var jobRunFinished = false
        var jobRunState: GlueClientTypes.JobRunState

        repeat {
            let jobRun = await getJobRun(glueClient: glueClient, name: jobName,
 id: jobRunID)
            guard let jobRun else {
                print("*** Unable to get the job run.")
                _ = await deleteJob(glueClient: glueClient, name: jobName)
                _ = await deleteDatabase(glueClient: glueClient, name:
 databaseName, withTables: true)
                _ = await deleteCrawler(glueClient: glueClient, name:
 crawlerName)
                return
            }
            jobRunState = jobRun.jobRunState ?? .failed

            //
=====================================================================
            // 7. Output where to find the data if the job run was successful.
            //    If the job run failed for any reason, output an appropriate
            //    error message.
            //
======================================================================

            switch jobRunState {
                case .succeeded:
                    print("Job run succeeded. JSON files are in the Amazon S3
 path:")
                    print("    \(outputURL)")
                    jobRunFinished = true
                case .stopped:
                    jobRunFinished = true
                case .error:
                    print("*** Error: Job run ended in an error.
 \(jobRun.errorMessage ?? "")")
                    jobRunFinished = true
                case .failed:
                    print("*** Error: Job run failed. \(jobRun.errorMessage ??
 "")")
                    jobRunFinished = true
                case .timeout:
                    print("*** Warning: Job run timed out.")
                    jobRunFinished = true
                default:
                    do {
                        try await Task.sleep(for: .milliseconds(250))
                    } catch {
                        print("*** Error pausing the task.")
                    }
            }
        } while jobRunFinished != true

        //====================================================================
        // 7.5. List the job runs for this job, showing each job run's ID and
        // its execution time.
        //====================================================================

        print("Getting all job runs for the job \(jobName):")
        let jobRuns = await getJobRuns(glueClient: glueClient, name: jobName)

        if jobRuns.count == 0 {
            print("    <no job runs found>")
        } else {
            print("Found \(jobRuns.count) job runs... listing execution times:")
            for jobRun in jobRuns {
                print("    \(jobRun.id ?? "<unnamed>"): \(jobRun.executionTime)
 seconds")
            }
        }

        //======================================================================
        // 8. List the jobs for the user's account.
        //======================================================================

        print("\nThe account has the following jobs:")
        let jobs = await listJobs(glueClient: glueClient)

        if jobs.count == 0 {
            print("    <no jobs found>")
        } else {
            for job in jobs {
                print("    \(job)")
            }
        }

        //======================================================================
        // 9. Get the job run details for a job run.
        //======================================================================

        print("Information about the job run:")
        let jobRun = await getJobRun(glueClient: glueClient, name: jobName, id:
 jobRunID)

        guard let jobRun else {
            print("*** Unable to retrieve the job run.")
            _ = await deleteJob(glueClient: glueClient, name: jobName)
            _ = await deleteDatabase(glueClient: glueClient, name: databaseName,
 withTables: true)
            _ = await deleteCrawler(glueClient: glueClient, name: crawlerName)
            return
        }

        let startDate = jobRun.startedOn ?? Date(timeIntervalSince1970: 0)
        let endDate = jobRun.completedOn ?? Date(timeIntervalSince1970: 0)
        let dateFormatter: DateFormatter = DateFormatter()
        dateFormatter.dateStyle = .long
        dateFormatter.timeStyle = .long
        print("    Started at: \(dateFormatter.string(from: startDate))")
        print("  Completed at: \(dateFormatter.string(from: endDate))")

        //======================================================================
        // 10. Delete the job.
        //======================================================================

        print("\nDeleting the job...")
        _ = await deleteJob(glueClient: glueClient, name: jobName)

        //======================================================================
        // 11. Delete the database and tables created by this example.
        //======================================================================

        print("Deleting the database...")
        _ = await deleteDatabase(glueClient: glueClient, name: databaseName,
 withTables: true)

        //======================================================================
        // 12. Delete the crawler.
        //======================================================================

        print("Deleting the crawler...")
        if !(await deleteCrawler(glueClient: glueClient, name: crawlerName)) {
            return
        }
    }
}

/// The program's asynchronous entry point.
@main
struct Main {
    static func main() async {
        let args = Array(CommandLine.arguments.dropFirst())

        do {
            let command = try ExampleCommand.parse(args)
            try await command.runAsync()
        } catch {
            ExampleCommand.exit(withError: error)
        }
    }
}
```

- For API details, see the following topics in *AWS SDK for Swift API reference*.
  - CreateCrawler
  - CreateJob
  - DeleteCrawler
  - DeleteDatabase
  - DeleteJob
  - DeleteTable
  - GetCrawler
  - GetDatabase
  - GetDatabases
  - GetJob
  - GetJobRun
  - GetJobRuns
  - GetTables
  - ListJobs
  - StartCrawler
  - StartJobRun

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Actions for AWS Glue using AWS SDKs

The following code examples demonstrate how to perform individual AWS Glue actions with AWS SDKs. Each example includes a link to GitHub, where you can find instructions for setting up and running the code.

The following examples include only the most commonly used actions. For a complete list, see the AWS Glue API Reference.

Examples

- Use CreateCrawler with an AWS SDK
- Use CreateJob with an AWS SDK or CLI
- Use DeleteCrawler with an AWS SDK
- Use DeleteDatabase with an AWS SDK
- Use DeleteJob with an AWS SDK or CLI
- Use DeleteTable with an AWS SDK
- Use GetCrawler with an AWS SDK
- Use GetDatabase with an AWS SDK
- Use GetDatabases with an AWS SDK or CLI
- Use GetJob with an AWS SDK or CLI
- Use GetJobRun with an AWS SDK or CLI
- Use GetJobRuns with an AWS SDK or CLI
- Use GetTables with an AWS SDK or CLI
- Use ListJobs with an AWS SDK
- Use StartCrawler with an AWS SDK or CLI
- Use StartJobRun with an AWS SDK or CLI

Use `CreateCrawler` with an AWS SDK

The following code examples show how to use `CreateCrawler`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Learn the basics

.NET

SDK for .NET

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
    /// <summary>
    /// Create an AWS Glue crawler.
    /// </summary>
    /// <param name="crawlerName">The name for the crawler.</param>
    /// <param name="crawlerDescription">A description of the crawler.</param>
    /// <param name="role">The AWS Identity and Access Management (IAM) role to
    /// be assumed by the crawler.</param>
    /// <param name="schedule">The schedule on which the crawler will be
 executed.</param>
    /// <param name="s3Path">The path to the Amazon Simple Storage Service
 (Amazon S3)
    /// bucket where the Python script has been stored.</param>
    /// <param name="dbName">The name to use for the database that will be
    /// created by the crawler.</param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> CreateCrawlerAsync(
        string crawlerName,
        string crawlerDescription,
        string role,
        string schedule,
        string s3Path,
        string dbName)
    {
        var s3Target = new S3Target
        {
            Path = s3Path,
        };

        var targetList = new List<S3Target>
        {
            s3Target,
        };

        var targets = new CrawlerTargets
        {
            S3Targets = targetList,
        };

        var crawlerRequest = new CreateCrawlerRequest
        {
            DatabaseName = dbName,
            Name = crawlerName,
            Description = crawlerDescription,
            Targets = targets,
            Role = role,
            Schedule = schedule,
        };

        var response = await _amazonGlue.CreateCrawlerAsync(crawlerRequest);
        return response.HttpStatusCode == System.Net.HttpStatusCode.OK;
    }
```

- For API details, see CreateCrawler in *AWS SDK for .NET API Reference*.

C++

SDK for C++

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region in which the bucket was created
 (overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::Glue::GlueClient client(clientConfig);

        Aws::Glue::Model::S3Target s3Target;
        s3Target.SetPath("s3://crawler-public-us-east-1/flight/2016/csv");
        Aws::Glue::Model::CrawlerTargets crawlerTargets;
        crawlerTargets.AddS3Targets(s3Target);

        Aws::Glue::Model::CreateCrawlerRequest request;
        request.SetTargets(crawlerTargets);
        request.SetName(CRAWLER_NAME);
        request.SetDatabaseName(CRAWLER_DATABASE_NAME);
        request.SetTablePrefix(CRAWLER_DATABASE_PREFIX);
        request.SetRole(roleArn);
        Aws::Glue::Model::CreateCrawlerOutcome outcome =
client.CreateCrawler(request);

        if (outcome.IsSuccess()) {
            std::cout << "Successfully created the crawler." << std::endl;
        }
        else {
            std::cerr << "Error creating a crawler. " <<
outcome.GetError().GetMessage()
                      << std::endl;
            deleteAssets("", CRAWLER_DATABASE_NAME, "", bucketName,
clientConfig);
            return false;
        }
```

- For API details, see CreateCrawler in *AWS SDK for C++ API Reference*.

Java

SDK for Java 2.x

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
    /**
     * Creates a new AWS Glue crawler using the AWS Glue Java API.
     *
     * @param glueClient  the AWS Glue client used to interact with the AWS Glue
service
     * @param iam         the IAM role that the crawler will use to access the
data source
     * @param s3Path      the S3 path that the crawler will scan for data
     * @param cron        the cron expression that defines the crawler's schedule
     * @param dbName      the name of the AWS Glue database where the crawler
will store the metadata
     * @param crawlerName the name of the crawler to be created
     */
    public static void createGlueCrawler(GlueClient glueClient,
                                         String iam,
                                         String s3Path,
                                         String cron,
                                         String dbName,
                                         String crawlerName) {

        try {
            S3Target s3Target = S3Target.builder()
                .path(s3Path)
                .build();

            List<S3Target> targetList = new ArrayList<>();
            targetList.add(s3Target);
            CrawlerTargets targets = CrawlerTargets.builder()
                .s3Targets(targetList)
                .build();

            CreateCrawlerRequest crawlerRequest = CreateCrawlerRequest.builder()
                .databaseName(dbName)
                .name(crawlerName)
                .description("Created by the AWS Glue Java API")
                .targets(targets)
                .role(iam)
                .schedule(cron)
                .build();

            glueClient.createCrawler(crawlerRequest);
            System.out.println(crawlerName + " was successfully created");

        } catch (GlueException e) {
            throw e;
        }
    }
```

- For API details, see CreateCrawler in *AWS SDK for Java 2.x API Reference*.

JavaScript

SDK for JavaScript (v3)

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
const createCrawler = (name, role, dbName, tablePrefix, s3TargetPath) => {
  const client = new GlueClient({});

  const command = new CreateCrawlerCommand({
    Name: name,
    Role: role,
    DatabaseName: dbName,
    TablePrefix: tablePrefix,
    Targets: {
      S3Targets: [{ Path: s3TargetPath }],
    },
  });

  return client.send(command);
};
```

- For API details, see CreateCrawler in *AWS SDK for JavaScript API Reference*.

Kotlin

SDK for Kotlin

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
suspend fun createGlueCrawler(
    iam: String?,
    s3Path: String?,
    cron: String?,
    dbName: String?,
    crawlerName: String,
) {
    val s3Target =
        S3Target {
            path = s3Path
        }

    // Add the S3Target to a list.
    val targetList = mutableListOf<S3Target>()
    targetList.add(s3Target)

    val targetOb =
        CrawlerTargets {
            s3Targets = targetList
        }

    val request =
        CreateCrawlerRequest {
            databaseName = dbName
            name = crawlerName
            description = "Created by the AWS Glue Kotlin API"
            targets = targetOb
            role = iam
            schedule = cron
        }

    GlueClient.fromEnvironment { region = "us-west-2" }.use { glueClient ->
        glueClient.createCrawler(request)
        println("$crawlerName was successfully created")
    }
}
```

- For API details, see CreateCrawler in *AWS SDK for Kotlin API reference*.

PHP

SDK for PHP

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        $crawlerName = "example-crawler-test-" . $uniqid;

        $role = $iamService->getRole("AWSGlueServiceRole-DocExample");

        $path = 's3://crawler-public-us-east-1/flight/2016/csv';
        $glueService->createCrawler($crawlerName, $role['Role']['Arn'],
$databaseName, $path);

    public function createCrawler($crawlerName, $role, $databaseName, $path):
Result
    {
        return $this->customWaiter(function () use ($crawlerName, $role,
$databaseName, $path) {
            return $this->glueClient->createCrawler([
                'Name' => $crawlerName,
                'Role' => $role,
                'DatabaseName' => $databaseName,
                'Targets' => [
                    'S3Targets' =>
                        [[
                            'Path' => $path,
                        ]]
                ],
            ]);
        });
    }
```

- For API details, see CreateCrawler in *AWS SDK for PHP API Reference*.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the [AWS Code Examples Repository](https://github.com/awsdocs/aws-doc-sdk-examples/tree/main/python/example_code/glue#code-examples).

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""

    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 Glue client.
        """
        self.glue_client = glue_client


    def create_crawler(self, name, role_arn, db_name, db_prefix, s3_target):
        """
        Creates a crawler that can crawl the specified target and populate a
        database in your AWS Glue Data Catalog with metadata that describes the
 data
        in the target.

        :param name: The name of the crawler.
        :param role_arn: The Amazon Resource Name (ARN) of an AWS Identity and
 Access
                         Management (IAM) role that grants permission to let AWS
 Glue
                         access the resources it needs.
        :param db_name: The name to give the database that is created by the
 crawler.
        :param db_prefix: The prefix to give any database tables that are created
 by
                          the crawler.
        :param s3_target: The URL to an S3 bucket that contains data that is
                          the target of the crawler.
        """
        try:
            self.glue_client.create_crawler(
                Name=name,
                Role=role_arn,
                DatabaseName=db_name,
                TablePrefix=db_prefix,
                Targets={"S3Targets": [{"Path": s3_target}]},
            )
        except ClientError as err:
            logger.error(
                "Couldn't create crawler. Here's why: %s: %s",
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise
```

- For API details, see CreateCrawler in *AWS SDK for Python (Boto3) API Reference*.

Ruby

SDK for Ruby

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
# The `GlueWrapper` class serves as a wrapper around the AWS Glue API, providing
 a simplified interface for common operations.
# It encapsulates the functionality of the AWS SDK for Glue and provides methods
 for interacting with Glue crawlers, databases, tables, jobs, and S3 resources.
# The class initializes with a Glue client and a logger, allowing it to make API
 calls and log any errors or informational messages.
class GlueWrapper
  def initialize(glue_client, logger)
    @glue_client = glue_client
    @logger = logger
  end
  # Creates a new crawler with the specified configuration.
  #
  # @param name [String] The name of the crawler.
  # @param role_arn [String] The ARN of the IAM role to be used by the crawler.
  # @param db_name [String] The name of the database where the crawler stores its
 metadata.
  # @param db_prefix [String] The prefix to be added to the names of tables that
 the crawler creates.
  # @param s3_target [String] The S3 path that the crawler will crawl.
  # @return [void]
  def create_crawler(name, role_arn, db_name, _db_prefix, s3_target)
    @glue_client.create_crawler(
      name: name,
      role: role_arn,
      database_name: db_name,
      targets: {
        s3_targets: [
          {
            path: s3_target
          }
        ]
      }
    )
  rescue Aws::Glue::Errors::GlueException => e
    @logger.error("Glue could not create crawler: \n#{e.message}")
    raise
  end
```

- For API details, see CreateCrawler in *AWS SDK for Ruby API Reference*.

Rust

SDK for Rust

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        let create_crawler = glue
```

```
            .create_crawler()
            .name(self.crawler())
            .database_name(self.database())
            .role(self.iam_role.expose_secret())
            .targets(
                CrawlerTargets::builder()
                    .s3_targets(S3Target::builder().path(CRAWLER_TARGET).build())
                    .build(),
            )
            .send()
            .await;

        match create_crawler {
            Err(err) => {
                let glue_err: aws_sdk_glue::Error = err.into();
                match glue_err {
                    aws_sdk_glue::Error::AlreadyExistsException(_) => {
                        info!("Using existing crawler");
                        Ok(())
                    }
                    _ => Err(GlueMvpError::GlueSdk(glue_err)),
                }
            }
            Ok(_) => Ok(()),
        }?;
```

- For API details, see CreateCrawler in *AWS SDK for Rust API reference*.

Swift

SDK for Swift

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
import AWSClientRuntime
import AWSGlue
```

```
   /// Create a new AWS Glue crawler.
   ///
   /// - Parameters:
   ///   - glueClient: An AWS Glue client to use for the crawler.
   ///   - crawlerName: A name for the new crawler.
   ///   - iamRole: The name of an Amazon IAM role for the crawler to use.
   ///   - s3Path: The path of an Amazon S3 folder to use as a target location.
   ///   - cronSchedule: A `cron` schedule indicating when to run the crawler.
   ///   - databaseName: The name of an AWS Glue database to operate on.
   ///
   /// - Returns: `true` if the crawler is created successfully, otherwise
`false`.
   func createCrawler(glueClient: GlueClient, crawlerName: String, iamRole:
String,
                      s3Path: String, cronSchedule: String, databaseName:
String) async -> Bool {
       let s3Target = GlueClientTypes.S3Target(path: s3url)
       let targetList = GlueClientTypes.CrawlerTargets(s3Targets: [s3Target])

       do {
           _ = try await glueClient.createCrawler(
               input: CreateCrawlerInput(
                   databaseName: databaseName,
                   description: "Created by the AWS SDK for Swift Scenario
Example for AWS Glue.",
                   name: crawlerName,
                   role: iamRole,
                   schedule: cronSchedule,
                   tablePrefix: tablePrefix,
                   targets: targetList
               )
           )
       } catch _ as AlreadyExistsException {
           print("*** A crawler named \"\(crawlerName)\" already exists.")
           return false
       } catch _ as OperationTimeoutException {
           print("*** The attempt to create the AWS Glue crawler timed out.")
           return false
       } catch {
           print("*** An unexpected error occurred creating the AWS Glue
crawler: \(error.localizedDescription)")
           return false
       }
```

```
            return true
        }
```

- For API details, see CreateCrawler in *AWS SDK for Swift API reference*.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `CreateJob` with an AWS SDK or CLI

The following code examples show how to use `CreateJob`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Learn the basics

.NET

SDK for .NET

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
    /// <summary>
    /// Create an AWS Glue job.
    /// </summary>
    /// <param name="jobName">The name of the job.</param>
    /// <param name="roleName">The name of the IAM role to be assumed by
    /// the job.</param>
    /// <param name="description">A description of the job.</param>
    /// <param name="scriptUrl">The URL to the script.</param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> CreateJobAsync(string dbName, string tableName,
string bucketUrl, string jobName, string roleName, string description, string
scriptUrl)
    {
        var command = new JobCommand
        {
            PythonVersion = "3",
            Name = "glueetl",
            ScriptLocation = scriptUrl,
        };

        var arguments = new Dictionary<string, string>
        {
            { "--input_database", dbName },
            { "--input_table", tableName },
            { "--output_bucket_url", bucketUrl }
        };

        var request = new CreateJobRequest
        {
            Command = command,
            DefaultArguments = arguments,
            Description = description,
            GlueVersion = "3.0",
            Name = jobName,
            NumberOfWorkers = 10,
            Role = roleName,
            WorkerType = "G.1X"
        };

        var response = await _amazonGlue.CreateJobAsync(request);
        return response.HttpStatusCode == HttpStatusCode.OK;
    }
```

- For API details, see CreateJob in *AWS SDK for .NET API Reference*.

C++

SDK for C++

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region in which the bucket was created
(overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::Glue::GlueClient client(clientConfig);

        Aws::Glue::Model::CreateJobRequest request;
        request.SetName(JOB_NAME);
        request.SetRole(roleArn);
        request.SetGlueVersion(GLUE_VERSION);

        Aws::Glue::Model::JobCommand command;
        command.SetName(JOB_COMMAND_NAME);
        command.SetPythonVersion(JOB_PYTHON_VERSION);
        command.SetScriptLocation(
                Aws::String("s3://") + bucketName + "/" + PYTHON_SCRIPT);
        request.SetCommand(command);

        Aws::Glue::Model::CreateJobOutcome outcome = client.CreateJob(request);

        if (outcome.IsSuccess()) {
            std::cout << "Successfully created the job." << std::endl;
        }
        else {
            std::cerr << "Error creating the job. " <<
outcome.GetError().GetMessage()
                      << std::endl;
            deleteAssets(CRAWLER_NAME, CRAWLER_DATABASE_NAME, "", bucketName,
                         clientConfig);
            return false;
        }
```

- For API details, see CreateJob in *AWS SDK for C++ API Reference*.

CLI

AWS CLI

To create a job to transform data

The following `create-job` example creates a streaming job that runs a script stored in S3.

```
aws glue create-job \
    --name my-testing-job \
    --role AWSGlueServiceRoleDefault \
    --command '{ \
        "Name": "gluestreaming", \
        "ScriptLocation": "s3://amzn-s3-demo-bucket/folder/" \
    }' \
    --region us-east-1 \
    --output json \
    --default-arguments '{ \
        "--job-language":"scala", \
        "--class":"GlueApp" \
    }' \
    --profile my-profile \
    --endpoint https://glue.us-east-1.amazonaws.com
```

Contents of `test_script.scala`:

```
import com.amazonaws.services.glue.ChoiceOption
import com.amazonaws.services.glue.GlueContext
import com.amazonaws.services.glue.MappingSpec
import com.amazonaws.services.glue.ResolveSpec
import com.amazonaws.services.glue.errors.CallSite
import com.amazonaws.services.glue.util.GlueArgParser
import com.amazonaws.services.glue.util.Job
import com.amazonaws.services.glue.util.JsonOptions
import org.apache.spark.SparkContext
import scala.collection.JavaConverters._

object GlueApp {
    def main(sysArgs: Array[String]) {
        val spark: SparkContext = new SparkContext()
        val glueContext: GlueContext = new GlueContext(spark)
        // @params: [JOB_NAME]
        val args = GlueArgParser.getResolvedOptions(sysArgs,
Seq("JOB_NAME").toArray)
        Job.init(args("JOB_NAME"), glueContext, args.asJava)
        // @type: DataSource
        // @args: [database = "tempdb", table_name = "s3-source",
transformation_ctx = "datasource0"]
        // @return: datasource0
        // @inputs: []
        val datasource0 = glueContext.getCatalogSource(database = "tempdb",
tableName = "s3-source", redshiftTmpDir = "", transformationContext =
"datasource0").getDynamicFrame()
        // @type: ApplyMapping
        // @args: [mapping = [("sensorid", "int", "sensorid", "int"),
("currenttemperature", "int", "currenttemperature", "int"), ("status", "string",
"status", "string")], transformation_ctx = "applymapping1"]
        // @return: applymapping1
        // @inputs: [frame = datasource0]
        val applymapping1 = datasource0.applyMapping(mappings = Seq(("sensorid",
"int", "sensorid", "int"), ("currenttemperature", "int", "currenttemperature",
"int"), ("status", "string", "status", "string")), caseSensitive = false,
transformationContext = "applymapping1")
        // @type: SelectFields
        // @args: [paths = ["sensorid", "currenttemperature", "status"],
transformation_ctx = "selectfields2"]
        // @return: selectfields2
        // @inputs: [frame = applymapping1]
        val selectfields2 = applymapping1.selectFields(paths = Seq("sensorid",
"currenttemperature", "status"), transformationContext = "selectfields2")
        // @type: ResolveChoice
        // @args: [choice = "MATCH_CATALOG", database = "tempdb", table_name =
"my-s3-sink", transformation_ctx = "resolvechoice3"]
        // @return: resolvechoice3
        // @inputs: [frame = selectfields2]
        val resolvechoice3 = selectfields2.resolveChoice(choiceOption =
Some(ChoiceOption("MATCH_CATALOG")), database = Some("tempdb"), tableName =
Some("my-s3-sink"), transformationContext = "resolvechoice3")
        // @type: DataSink
        // @args: [database = "tempdb", table_name = "my-s3-sink",
transformation_ctx = "datasink4"]
        // @return: datasink4
        // @inputs: [frame = resolvechoice3]
        val datasink4 = glueContext.getCatalogSink(database = "tempdb",
 tableName = "my-s3-sink", redshiftTmpDir = "", transformationContext =
 "datasink4").writeDynamicFrame(resolvechoice3)
        Job.commit()
    }
}
```

Output:

```
{
    "Name": "my-testing-job"
}
```

For more information, see Authoring Jobs in AWS Glue in the *AWS Glue Developer Guide*.

- For API details, see CreateJob in *AWS CLI Command Reference*.

Java

SDK for Java 2.x

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```

    /**
     * Creates a new AWS Glue job.
     *
     * @param glueClient     the AWS Glue client to use for the operation
     * @param jobName        the name of the job to create
     * @param iam            the IAM role to associate with the job
     * @param scriptLocation the location of the script to be used by the job
     * @throws GlueException if there is an error creating the job
     */
    public static void createJob(GlueClient glueClient, String jobName, String
 iam, String scriptLocation) {
        try {
            JobCommand command = JobCommand.builder()
```

```
                .pythonVersion("3")
                .name("glueetl")
                .scriptLocation(scriptLocation)
                .build();

            CreateJobRequest jobRequest = CreateJobRequest.builder()
                .description("A Job created by using the AWS SDK for Java V2")
                .glueVersion("2.0")
                .workerType(WorkerType.G_1_X)
                .numberOfWorkers(10)
                .name(jobName)
                .role(iam)
                .command(command)
                .build();

            glueClient.createJob(jobRequest);
            System.out.println(jobName + " was successfully created.");

        } catch (GlueException e) {
            throw e;
        }
    }
```

- For API details, see CreateJob in *AWS SDK for Java 2.x API Reference*.

JavaScript

SDK for JavaScript (v3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
const createJob = (name, role, scriptBucketName, scriptKey) => {
  const client = new GlueClient({});

  const command = new CreateJobCommand({
    Name: name,
    Role: role,
    Command: {
      Name: "glueetl",
      PythonVersion: "3",
      ScriptLocation: `s3://${scriptBucketName}/${scriptKey}`,
    },
    GlueVersion: "3.0",
  });

  return client.send(command);
};
```

- For API details, see CreateJob in *AWS SDK for JavaScript API Reference*.

PHP

SDK for PHP

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        $role = $iamService->getRole("AWSGlueServiceRole-DocExample");

        $jobName = 'test-job-' . $uniqid;

        $scriptLocation = "s3://$bucketName/run_job.py";
        $job = $glueService->createJob($jobName, $role['Role']['Arn'],
 $scriptLocation);

    public function createJob($jobName, $role, $scriptLocation, $pythonVersion =
 '3', $glueVersion = '3.0'): Result
    {
        return $this->glueClient->createJob([
            'Name' => $jobName,
            'Role' => $role,
            'Command' => [
                'Name' => 'glueetl',
                'ScriptLocation' => $scriptLocation,
                'PythonVersion' => $pythonVersion,
            ],
            'GlueVersion' => $glueVersion,
        ]);
    }
```

- For API details, see [CreateJob](#) in *AWS SDK for PHP API Reference*.

PowerShell

Tools for PowerShell V4

Example 1: This example creates a new job in AWS Glue. The command name value is always `glueetl`. AWS Glue supports running job scripts written in Python or Scala. In this example, the job script (MyTestGlueJob.py) is written in Python. Python parameters are specified in the `$DefArgs` variable, and then passed to the PowerShell command in the `DefaultArguments` parameter, which accepts a hashtable. The parameters in the `$JobParams` variable come from the CreateJob API, documented in the Jobs (https://docs.aws.amazon.com/glue/latest/dg/aws-glue-api-jobs-job.html) topic of the AWS Glue API reference.

```
$Command = New-Object Amazon.Glue.Model.JobCommand
$Command.Name = 'glueetl'
$Command.ScriptLocation = 's3://amzn-s3-demo-source-bucket/admin/
MyTestGlueJob.py'
$Command

$Source = "source_test_table"
$Target = "target_test_table"
$Connections = $Source, $Target

$DefArgs = @{
     '--TempDir' = 's3://amzn-s3-demo-bucket/admin'
     '--job-bookmark-option' = 'job-bookmark-disable'
     '--job-language' = 'python'
     }
$DefArgs

$ExecutionProp = New-Object Amazon.Glue.Model.ExecutionProperty
$ExecutionProp.MaxConcurrentRuns = 1
$ExecutionProp
```

```
$JobParams = @{
    "AllocatedCapacity"    = "5"
    "Command"              = $Command
    "Connections_Connection" = $Connections
    "DefaultArguments"  = $DefArgs
    "Description"       = "This is a test"
    "ExecutionProperty" = $ExecutionProp
    "MaxRetries"        = "1"
    "Name"              = "MyOregonTestGlueJob"
    "Role"              = "Amazon-GlueServiceRoleForSSM"
    "Timeout"           = "20"
     }

New-GlueJob @JobParams
```

- For API details, see CreateJob in *AWS Tools for PowerShell Cmdlet Reference (V4)*.

Tools for PowerShell V5

Example 1: This example creates a new job in AWS Glue. The command name value is always `glueetl`. AWS Glue supports running job scripts written in Python or Scala. In this example, the job script (MyTestGlueJob.py) is written in Python. Python parameters are specified in the `$DefArgs` variable, and then passed to the PowerShell command in the `DefaultArguments` parameter, which accepts a hashtable. The parameters in the `$JobParams` variable come from the CreateJob API, documented in the Jobs (https://docs.aws.amazon.com/glue/latest/dg/aws-glue-api-jobs-job.html) topic of the AWS Glue API reference.

```
$Command = New-Object Amazon.Glue.Model.JobCommand
$Command.Name = 'glueetl'
$Command.ScriptLocation = 's3://amzn-s3-demo-source-bucket/admin/
MyTestGlueJob.py'
$Command

$Source = "source_test_table"
$Target = "target_test_table"
$Connections = $Source, $Target

$DefArgs = @{
     '--TempDir' = 's3://amzn-s3-demo-bucket/admin'
     '--job-bookmark-option' = 'job-bookmark-disable'
     '--job-language' = 'python'
     }
$DefArgs

$ExecutionProp = New-Object Amazon.Glue.Model.ExecutionProperty
$ExecutionProp.MaxConcurrentRuns = 1
$ExecutionProp

$JobParams = @{
    "AllocatedCapacity"    = "5"
    "Command"              = $Command
    "Connections_Connection" = $Connections
    "DefaultArguments"  = $DefArgs
    "Description"       = "This is a test"
    "ExecutionProperty" = $ExecutionProp
    "MaxRetries"        = "1"
    "Name"              = "MyOregonTestGlueJob"
    "Role"              = "Amazon-GlueServiceRoleForSSM"
    "Timeout"           = "20"
     }

New-GlueJob @JobParams
```

- For API details, see CreateJob in *AWS Tools for PowerShell Cmdlet Reference (V5)*.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""

    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 Glue client.
        """
        self.glue_client = glue_client
    def create_job(self, name, description, role_arn, script_location):
        """
        Creates a job definition for an extract, transform, and load (ETL) job 
that can
        be run by AWS Glue.

        :param name: The name of the job definition.
        :param description: The description of the job definition.
        :param role_arn: The ARN of an IAM role that grants AWS Glue the 
permissions
                         it requires to run the job.
        :param script_location: The Amazon S3 URL of a Python ETL script that is 
run as
                                part of the job. The script defines how the data 
is
                                transformed.
        """
        try:
            self.glue_client.create_job(
                Name=name,
                Description=description,
                Role=role_arn,
                Command={
                    "Name": "glueetl",
                    "ScriptLocation": script_location,
                    "PythonVersion": "3",
                },
                GlueVersion="3.0",
            )
        except ClientError as err:
            logger.error(
                "Couldn't create job %s. Here's why: %s: %s",
                name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise
```

- For API details, see CreateJob in *AWS SDK for Python (Boto3) API Reference*.

Ruby

SDK for Ruby

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
# The `GlueWrapper` class serves as a wrapper around the AWS Glue API, providing
 a simplified interface for common operations.
# It encapsulates the functionality of the AWS SDK for Glue and provides methods
 for interacting with Glue crawlers, databases, tables, jobs, and S3 resources.
# The class initializes with a Glue client and a logger, allowing it to make API
 calls and log any errors or informational messages.
class GlueWrapper
  def initialize(glue_client, logger)
    @glue_client = glue_client
    @logger = logger
  end

  # Creates a new job with the specified configuration.
  #
  # @param name [String] The name of the job.
  # @param description [String] The description of the job.
  # @param role_arn [String] The ARN of the IAM role to be used by the job.
  # @param script_location [String] The location of the ETL script for the job.
  # @return [void]
  def create_job(name, description, role_arn, script_location)
    @glue_client.create_job(
      name: name,
      description: description,
      role: role_arn,
      command: {
        name: 'glueetl',
        script_location: script_location,
        python_version: '3'
      },
      glue_version: '3.0'
    )
```

```
rescue Aws::Glue::Errors::GlueException => e
  @logger.error("Glue could not create job #{name}: \n#{e.message}")
  raise
end
```

- For API details, see CreateJob in *AWS SDK for Ruby API Reference*.

Rust

SDK for Rust

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        let create_job = glue
            .create_job()
            .name(self.job())
            .role(self.iam_role.expose_secret())
            .command(
                JobCommand::builder()
                    .name("glueetl")
                    .python_version("3")
                    .script_location(format!("s3://{}/job.py", self.bucket()))
                    .build(),
            )
            .glue_version("3.0")
            .send()
            .await
            .map_err(GlueMvpError::from_glue_sdk)?;

        let job_name = create_job.name().ok_or_else(|| {
            GlueMvpError::Unknown("Did not get job name after creating
job".into())
        })?;
```

- For API details, see CreateJob in *AWS SDK for Rust API reference*.

Swift

SDK for Swift

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the [AWS Code Examples Repository](https://github.com/awsdocs/aws-doc-sdk-examples).

```
import AWSClientRuntime
import AWSGlue

    /// Create a new AWS Glue job.
    ///
    /// - Parameters:
    ///   - glueClient: The AWS Glue client to use.
    ///   - jobName: The name to give the new job.
    ///   - role: The IAM role for the job to use when accessing AWS services.
    ///   - scriptLocation: The AWS S3 URI of the script to be run by the job.
    ///
    /// - Returns: `true` if the job is created successfully, otherwise `false`.
    func createJob(glueClient: GlueClient, name jobName: String, role: String,
                   scriptLocation: String) async -> Bool {
        let command = GlueClientTypes.JobCommand(
            name: "glueetl",
            pythonVersion: "3",
            scriptLocation: scriptLocation
        )

        do {
            _ = try await glueClient.createJob(
                input: CreateJobInput(
                    command: command,
                    description: "Created by the AWS SDK for Swift Glue basic
 scenario example.",
                    glueVersion: "3.0",
                    name: jobName,
                    numberOfWorkers: 10,
                    role: role,
                    workerType: .g1x
                )
            )
        } catch {
            return false
        }
        return true
    }
```

- For API details, see CreateJob in *AWS SDK for Swift API reference*.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `DeleteCrawler` with an AWS SDK

The following code examples show how to use `DeleteCrawler`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Learn the basics

.NET

SDK for .NET

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
    /// <summary>
    /// Delete an AWS Glue crawler.
    /// </summary>
    /// <param name="crawlerName">The name of the crawler.</param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> DeleteCrawlerAsync(string crawlerName)
    {
        var response = await _amazonGlue.DeleteCrawlerAsync(new
DeleteCrawlerRequest { Name = crawlerName });
        return response.HttpStatusCode == HttpStatusCode.OK;
    }
```

- For API details, see DeleteCrawler in *AWS SDK for .NET API Reference*.

C++

SDK for C++

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region in which the bucket was created
(overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::Glue::GlueClient client(clientConfig);

        Aws::Glue::Model::DeleteCrawlerRequest request;
        request.SetName(crawler);

        Aws::Glue::Model::DeleteCrawlerOutcome outcome =
client.DeleteCrawler(request);

        if (outcome.IsSuccess()) {
            std::cout << "Successfully deleted the crawler." << std::endl;
        }
        else {
            std::cerr << "Error deleting the crawler. "
                      << outcome.GetError().GetMessage() << std::endl;
            result = false;
        }
```

- For API details, see DeleteCrawler in *AWS SDK for C++ API Reference*.

Java

SDK for Java 2.x

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
    /**
     * Deletes a specific AWS Glue crawler.
     *
     * @param glueClient  the AWS Glue client object
     * @param crawlerName the name of the crawler to be deleted
     * @throws GlueException if an error occurs during the deletion process
     */
    public static void deleteSpecificCrawler(GlueClient glueClient, String
 crawlerName) {
        try {
            DeleteCrawlerRequest deleteCrawlerRequest =
 DeleteCrawlerRequest.builder()
                .name(crawlerName)
                .build();

            glueClient.deleteCrawler(deleteCrawlerRequest);
            System.out.println(crawlerName + " was deleted");

        } catch (GlueException e) {
            throw e;
        }
    }
```

- For API details, see DeleteCrawler in *AWS SDK for Java 2.x API Reference*.

JavaScript

SDK for JavaScript (v3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
const deleteCrawler = (crawlerName) => {
  const client = new GlueClient({});

  const command = new DeleteCrawlerCommand({
    Name: crawlerName,
  });

  return client.send(command);
};
```

- For API details, see DeleteCrawler in *AWS SDK for JavaScript API Reference*.

PHP

SDK for PHP

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        echo "Delete the crawler.\n";
        $glueClient->deleteCrawler([
            'Name' => $crawlerName,
        ]);

    public function deleteCrawler($crawlerName)
    {
        return $this->glueClient->deleteCrawler([
            'Name' => $crawlerName,
        ]);
    }
```

- For API details, see DeleteCrawler in *AWS SDK for PHP API Reference*.

Python

SDK for Python (Boto3)

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""

    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 Glue client.
        """
        self.glue_client = glue_client


    def delete_crawler(self, name):
        """
        Deletes a crawler.

        :param name: The name of the crawler to delete.
        """
        try:
            self.glue_client.delete_crawler(Name=name)
        except ClientError as err:
            logger.error(
                "Couldn't delete crawler %s. Here's why: %s: %s",
                name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise
```

- For API details, see DeleteCrawler in *AWS SDK for Python (Boto3) API Reference*.

Ruby

SDK for Ruby

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
# The `GlueWrapper` class serves as a wrapper around the AWS Glue API, providing
 a simplified interface for common operations.
# It encapsulates the functionality of the AWS SDK for Glue and provides methods
 for interacting with Glue crawlers, databases, tables, jobs, and S3 resources.
# The class initializes with a Glue client and a logger, allowing it to make API
 calls and log any errors or informational messages.
class GlueWrapper
  def initialize(glue_client, logger)
    @glue_client = glue_client
    @logger = logger
  end

  # Deletes a crawler with the specified name.
  #
  # @param name [String] The name of the crawler to delete.
  # @return [void]
  def delete_crawler(name)
    @glue_client.delete_crawler(name: name)
  rescue Aws::Glue::Errors::ServiceError => e
    @logger.error("Glue could not delete crawler #{name}: \n#{e.message}")
    raise
  end
```

- For API details, see DeleteCrawler in *AWS SDK for Ruby API Reference*.

Rust

SDK for Rust

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        glue.delete_crawler()
            .name(self.crawler())
            .send()
            .await
            .map_err(GlueMvpError::from_glue_sdk)?;
```

- For API details, see DeleteCrawler in *AWS SDK for Rust API reference*.

Swift

SDK for Swift

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
import AWSClientRuntime
import AWSGlue

    /// Delete an AWS Glue crawler.
    ///
    /// - Parameters:
    ///   - glueClient: The AWS Glue client to use.
    ///   - name: The name of the crawler to delete.
    ///
    /// - Returns: `true` if successful, otherwise `false`.
    func deleteCrawler(glueClient: GlueClient, name: String) async -> Bool {
        do {
            _ = try await glueClient.deleteCrawler(
                input: DeleteCrawlerInput(name: name)
            )
        } catch {
            return false
        }
        return true
    }
```

- For API details, see DeleteCrawler in *AWS SDK for Swift API reference*.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `DeleteDatabase` with an AWS SDK

The following code examples show how to use `DeleteDatabase`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Learn the basics

.NET

SDK for .NET

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
    /// <summary>
    /// Delete the AWS Glue database.
    /// </summary>
    /// <param name="dbName">The name of the database.</param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> DeleteDatabaseAsync(string dbName)
    {
        var response = await _amazonGlue.DeleteDatabaseAsync(new
 DeleteDatabaseRequest { Name = dbName });
        return response.HttpStatusCode == HttpStatusCode.OK;
    }
```

- For API details, see DeleteDatabase in *AWS SDK for .NET API Reference*.

C++

SDK for C++

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region in which the bucket was created
 (overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::Glue::GlueClient client(clientConfig);

        Aws::Glue::Model::DeleteDatabaseRequest request;
        request.SetName(database);

        Aws::Glue::Model::DeleteDatabaseOutcome outcome = client.DeleteDatabase(
                request);

        if (outcome.IsSuccess()) {
            std::cout << "Successfully deleted the database." << std::endl;
        }
        else {
            std::cerr << "Error deleting database. " <<
outcome.GetError().GetMessage()
                      << std::endl;
            result = false;
        }
```

- For API details, see DeleteDatabase in *AWS SDK for C++ API Reference*.

Java

SDK for Java 2.x

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
    /**
     * Deletes a AWS Glue Database.
     *
     * @param glueClient   An instance of the AWS Glue client used to interact
with the AWS Glue service.
     * @param databaseName The name of the database to be deleted.
     * @throws GlueException If an error occurs while deleting the database.
     */
    public static void deleteDatabase(GlueClient glueClient, String databaseName)
{
        try {
            DeleteDatabaseRequest request = DeleteDatabaseRequest.builder()
                .name(databaseName)
                .build();

            glueClient.deleteDatabase(request);
            System.out.println(databaseName + " was successfully deleted");

        } catch (GlueException e) {
            throw e;
        }
    }
```

- For API details, see DeleteDatabase in *AWS SDK for Java 2.x API Reference*.

JavaScript

SDK for JavaScript (v3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
const deleteDatabase = (databaseName) => {
  const client = new GlueClient({});

  const command = new DeleteDatabaseCommand({
    Name: databaseName,
  });

  return client.send(command);
};
```

- For API details, see DeleteDatabase in *AWS SDK for JavaScript API Reference*.

PHP

SDK for PHP

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        echo "Delete the databases.\n";
        $glueClient->deleteDatabase([
            'Name' => $databaseName,
        ]);

    public function deleteDatabase($databaseName)
    {
        return $this->glueClient->deleteDatabase([
            'Name' => $databaseName,
        ]);
    }
```

- For API details, see DeleteDatabase in *AWS SDK for PHP API Reference*.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""

    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 Glue client.
        """
        self.glue_client = glue_client


    def delete_database(self, name):
        """
        Deletes a metadata database from your Data Catalog.

        :param name: The name of the database to delete.
        """
        try:
            self.glue_client.delete_database(Name=name)
        except ClientError as err:
            logger.error(
                "Couldn't delete database %s. Here's why: %s: %s",
                name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise
```

- For API details, see DeleteDatabase in *AWS SDK for Python (Boto3) API Reference*.

Ruby

SDK for Ruby

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
# The `GlueWrapper` class serves as a wrapper around the AWS Glue API, providing
 a simplified interface for common operations.
# It encapsulates the functionality of the AWS SDK for Glue and provides methods
 for interacting with Glue crawlers, databases, tables, jobs, and S3 resources.
# The class initializes with a Glue client and a logger, allowing it to make API
 calls and log any errors or informational messages.
class GlueWrapper
  def initialize(glue_client, logger)
    @glue_client = glue_client
    @logger = logger
  end

  # Removes a specified database from a Data Catalog.
  #
  # @param database_name [String] The name of the database to delete.
  # @return [void]
  def delete_database(database_name)
    @glue_client.delete_database(name: database_name)
  rescue Aws::Glue::Errors::ServiceError => e
```

```
    @logger.error("Glue could not delete database: \n#{e.message}")
  end
```

- For API details, see DeleteDatabase in *AWS SDK for Ruby API Reference*.

Rust

SDK for Rust

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        glue.delete_database()
            .name(self.database())
            .send()
            .await
            .map_err(GlueMvpError::from_glue_sdk)?;
```

- For API details, see DeleteDatabase in *AWS SDK for Rust API reference*.

Swift

SDK for Swift

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
import AWSClientRuntime
import AWSGlue

    /// Delete the specified database.
    ///
    /// - Parameters:
    ///   - glueClient: The AWS Glue client to use.
    ///   - databaseName: The name of the database to delete.
    ///   - deleteTables: A Bool indicating whether or not to delete the
    ///     tables in the database before attempting to delete the database.
    ///
    /// - Returns: `true` if the database (and optionally its tables) are
    ///   deleted, otherwise `false`.
    func deleteDatabase(glueClient: GlueClient, name databaseName: String,
                        withTables deleteTables: Bool = false) async -> Bool {
        if deleteTables {
            var tableNames: [String] = []

            // Get a list of the names of all of the tables in the database.

            let tableList = await self.getTablesInDatabase(glueClient:
 glueClient, databaseName: databaseName)
            for table in tableList {
                guard let name = table.name else {
                    continue
                }
                tableNames.append(name)
            }

            // Delete the tables. If there's only one table, use
            // `deleteTable()`, otherwise, use `batchDeleteTable()`. You can
            // use `batchDeleteTable()` for a single table, but this
            // demonstrates the use of `deleteTable()`.

            if tableNames.count == 1 {
                do {
                    print("    Deleting table...")
                    _ = try await glueClient.deleteTable(
                        input: DeleteTableInput(
                            databaseName: databaseName,
                            name: tableNames[0]
                        )
                    )
                } catch {
                    print("*** Unable to delete the table.")
                }
            } else {
                do {
                    print("    Deleting tables...")
                    _ = try await glueClient.batchDeleteTable(
                        input: BatchDeleteTableInput(
                            databaseName: databaseName,
                            tablesToDelete: tableNames
                        )
                    )
                } catch {
                    print("*** Unable to delete the tables.")
                }
            }
        }

        // Delete the database itself.

        do {
            print("    Deleting the database itself...")
            _ = try await glueClient.deleteDatabase(
                input: DeleteDatabaseInput(name: databaseName)
            )
        } catch {
            print("*** Unable to delete the database.")
            return false
        }
        return true
    }
```

- For API details, see DeleteDatabase in *AWS SDK for Swift API reference*.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `DeleteJob` with an AWS SDK or CLI

The following code examples show how to use `DeleteJob`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Learn the basics

.NET

SDK for .NET

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the [AWS Code Examples Repository](https://github.com/awsdocs/aws-doc-sdk-examples/tree/main/dotnetv3/Glue#code-examples).

```
    /// <summary>
    /// Delete an AWS Glue job.
    /// </summary>
    /// <param name="jobName">The name of the job.</param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> DeleteJobAsync(string jobName)
    {
        var response = await _amazonGlue.DeleteJobAsync(new DeleteJobRequest
 { JobName = jobName });
        return response.HttpStatusCode == HttpStatusCode.OK;
    }
```

- For API details, see DeleteJob in *AWS SDK for .NET API Reference*.

C++

SDK for C++

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region in which the bucket was created
 (overrides config file).
        // clientConfig.region = "us-east-1";
```

```
    Aws::Glue::GlueClient client(clientConfig);

        Aws::Glue::Model::DeleteJobRequest request;
        request.SetJobName(job);

        Aws::Glue::Model::DeleteJobOutcome outcome = client.DeleteJob(request);


        if (outcome.IsSuccess()) {
            std::cout << "Successfully deleted the job." << std::endl;
        }
        else {
            std::cerr << "Error deleting the job. " <<
 outcome.GetError().GetMessage()
                      << std::endl;
            result = false;
        }
```

- For API details, see DeleteJob in *AWS SDK for C++ API Reference*.

CLI

AWS CLI

To delete a job

The following `delete-job` example deletes a job that is no longer needed.

```
aws glue delete-job \
    --job-name my-testing-job
```

Output:

```
{
    "JobName": "my-testing-job"
}
```

For more information, see Working with Jobs on the AWS Glue Console in the *AWS Glue Developer Guide*.

- For API details, see DeleteJob in *AWS CLI Command Reference*.

Java

SDK for Java 2.x

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
    /**
     * Deletes a Glue job.
     *
     * @param glueClient the Glue client to use for the operation
     * @param jobName    the name of the job to be deleted
     * @throws GlueException if there is an error deleting the job
     */
    public static void deleteJob(GlueClient glueClient, String jobName) {
        try {
            DeleteJobRequest jobRequest = DeleteJobRequest.builder()
                .jobName(jobName)
                .build();

            glueClient.deleteJob(jobRequest);
            System.out.println(jobName + " was successfully deleted");

        } catch (GlueException e) {
            throw e;
        }
    }
```

- For API details, see DeleteJob in *AWS SDK for Java 2.x API Reference*.

JavaScript

SDK for JavaScript (v3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
const deleteJob = (jobName) => {
  const client = new GlueClient({});

  const command = new DeleteJobCommand({
    JobName: jobName,
  });

  return client.send(command);
};
```

- For API details, see DeleteJob in *AWS SDK for JavaScript API Reference*.

PHP

SDK for PHP

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        echo "Delete the job.\n";
        $glueClient->deleteJob([
            'JobName' => $job['Name'],
        ]);

    public function deleteJob($jobName)
    {
        return $this->glueClient->deleteJob([
            'JobName' => $jobName,
        ]);
    }
```

- For API details, see DeleteJob in *AWS SDK for PHP API Reference*.

Python

SDK for Python (Boto3)

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""

    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 Glue client.
        """
        self.glue_client = glue_client


    def delete_job(self, job_name):
        """
        Deletes a job definition. This also deletes data about all runs that are
        associated with this job definition.

        :param job_name: The name of the job definition to delete.
        """
        try:
            self.glue_client.delete_job(JobName=job_name)
        except ClientError as err:
            logger.error(
                "Couldn't delete job %s. Here's why: %s: %s",
                job_name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise
```

- For API details, see DeleteJob in *AWS SDK for Python (Boto3) API Reference*.

Ruby

SDK for Ruby

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
# The `GlueWrapper` class serves as a wrapper around the AWS Glue API, providing
 a simplified interface for common operations.
# It encapsulates the functionality of the AWS SDK for Glue and provides methods
 for interacting with Glue crawlers, databases, tables, jobs, and S3 resources.
# The class initializes with a Glue client and a logger, allowing it to make API
 calls and log any errors or informational messages.
class GlueWrapper
  def initialize(glue_client, logger)
    @glue_client = glue_client
    @logger = logger
  end

  # Deletes a job with the specified name.
  #
  # @param job_name [String] The name of the job to delete.
  # @return [void]
  def delete_job(job_name)
    @glue_client.delete_job(job_name: job_name)
  rescue Aws::Glue::Errors::ServiceError => e
    @logger.error("Glue could not delete job: \n#{e.message}")
  end
```

- For API details, see DeleteJob in *AWS SDK for Ruby API Reference*.

Rust

SDK for Rust

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        glue.delete_job()
            .job_name(self.job())
            .send()
            .await
            .map_err(GlueMvpError::from_glue_sdk)?;
```

- For API details, see DeleteJob in *AWS SDK for Rust API reference*.

Swift

SDK for Swift

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
import AWSClientRuntime
import AWSGlue

    /// Delete an AWS Glue job.
    ///
    /// - Parameters:
    ///   - glueClient: The AWS Glue client to use.
    ///   - jobName: The name of the job to delete.
    ///
    /// - Returns: `true` if the job is successfully deleted, otherwise `false`.
    func deleteJob(glueClient: GlueClient, name jobName: String) async -> Bool {
        do {
            _ = try await glueClient.deleteJob(
                input: DeleteJobInput(jobName: jobName)
            )
        } catch {
            return false
        }
        return true
    }
```

- For API details, see DeleteJob in *AWS SDK for Swift API reference*.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `DeleteTable` with an AWS SDK

The following code examples show how to use `DeleteTable`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Learn the basics

.NET

SDK for .NET

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
    /// <summary>
    /// Delete a table from an AWS Glue database.
    /// </summary>
    /// <param name="tableName">The table to delete.</param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> DeleteTableAsync(string dbName, string tableName)
    {
        var response = await _amazonGlue.DeleteTableAsync(new DeleteTableRequest
 { Name = tableName, DatabaseName = dbName });
        return response.HttpStatusCode == HttpStatusCode.OK;
    }
```

- For API details, see DeleteTable in *AWS SDK for .NET API Reference*.

JavaScript

SDK for JavaScript (v3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
const deleteTable = (databaseName, tableName) => {
  const client = new GlueClient({});

  const command = new DeleteTableCommand({
    DatabaseName: databaseName,
    Name: tableName,
  });

  return client.send(command);
};
```

- For API details, see DeleteTable in *AWS SDK for JavaScript API Reference*.

PHP

SDK for PHP

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        echo "Delete the tables.\n";
        foreach ($tables['TableList'] as $table) {
            $glueService->deleteTable($table['Name'], $databaseName);
        }

    public function deleteTable($tableName, $databaseName)
    {
        return $this->glueClient->deleteTable([
            'DatabaseName' => $databaseName,
            'Name' => $tableName,
        ]);
    }
```

- For API details, see DeleteTable in *AWS SDK for PHP API Reference*.

Python

SDK for Python (Boto3)

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""

    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 Glue client.
        """
        self.glue_client = glue_client


    def delete_table(self, db_name, table_name):
        """
        Deletes a table from a metadata database.

        :param db_name: The name of the database that contains the table.
        :param table_name: The name of the table to delete.
        """
        try:
            self.glue_client.delete_table(DatabaseName=db_name, Name=table_name)
        except ClientError as err:
            logger.error(
                "Couldn't delete table %s. Here's why: %s: %s",
                table_name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise
```

- For API details, see [DeleteTable](#) in *AWS SDK for Python (Boto3) API Reference*.

Ruby

SDK for Ruby

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the [AWS Code Examples Repository](#).

```
# The `GlueWrapper` class serves as a wrapper around the AWS Glue API, providing
 a simplified interface for common operations.
```

```
# It encapsulates the functionality of the AWS SDK for Glue and provides methods
 for interacting with Glue crawlers, databases, tables, jobs, and S3 resources.
# The class initializes with a Glue client and a logger, allowing it to make API
 calls and log any errors or informational messages.
class GlueWrapper
  def initialize(glue_client, logger)
    @glue_client = glue_client
    @logger = logger
  end

  # Deletes a table with the specified name.
  #
  # @param database_name [String] The name of the catalog database in which the
 table resides.
  # @param table_name [String] The name of the table to be deleted.
  # @return [void]
  def delete_table(database_name, table_name)
    @glue_client.delete_table(database_name: database_name, name: table_name)
  rescue Aws::Glue::Errors::ServiceError => e
    @logger.error("Glue could not delete job: \n#{e.message}")
  end
```

- For API details, see DeleteTable in *AWS SDK for Ruby API Reference*.

Rust

SDK for Rust

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        for t in &self.tables {
            glue.delete_table()
                .name(t.name())
                .database_name(self.database())
                .send()
                .await
                .map_err(GlueMvpError::from_glue_sdk)?;
        }
```

- For API details, see DeleteTable in *AWS SDK for Rust API reference*.

Swift

SDK for Swift

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
import AWSClientRuntime
import AWSGlue

                do {
                    print("    Deleting table...")
                    _ = try await glueClient.deleteTable(
                        input: DeleteTableInput(
                            databaseName: databaseName,
                            name: tableNames[0]
                        )
                    )
                } catch {
                    print("*** Unable to delete the table.")
                }
```

- For API details, see DeleteTable in *AWS SDK for Swift API reference*.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `GetCrawler` with an AWS SDK

The following code examples show how to use `GetCrawler`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Learn the basics

.NET

SDK for .NET

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
    /// <summary>
    /// Get information about an AWS Glue crawler.
    /// </summary>
    /// <param name="crawlerName">The name of the crawler.</param>
    /// <returns>A Crawler object describing the crawler.</returns>
    public async Task<Crawler?> GetCrawlerAsync(string crawlerName)
    {
        var crawlerRequest = new GetCrawlerRequest
        {
            Name = crawlerName,
        };

        var response = await _amazonGlue.GetCrawlerAsync(crawlerRequest);
        if (response.HttpStatusCode == System.Net.HttpStatusCode.OK)
        {
            var databaseName = response.Crawler.DatabaseName;
            Console.WriteLine($"{crawlerName} has the database {databaseName}");
            return response.Crawler;
        }

        Console.WriteLine($"No information regarding {crawlerName} could be
 found.");
        return null;
    }
```

- For API details, see GetCrawler in *AWS SDK for .NET API Reference*.

C++

SDK for C++

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region in which the bucket was created
(overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::Glue::GlueClient client(clientConfig);

        Aws::Glue::Model::GetCrawlerRequest request;
        request.SetName(CRAWLER_NAME);

        Aws::Glue::Model::GetCrawlerOutcome outcome = client.GetCrawler(request);

        if (outcome.IsSuccess()) {
            Aws::Glue::Model::CrawlerState crawlerState =
outcome.GetResult().GetCrawler().GetState();
            std::cout << "Retrieved crawler with state " <<

Aws::Glue::Model::CrawlerStateMapper::GetNameForCrawlerState(
                             crawlerState)
                      << "." << std::endl;
        }
        else {
            std::cerr << "Error retrieving a crawler.  "
                      << outcome.GetError().GetMessage() << std::endl;
            deleteAssets(CRAWLER_NAME, CRAWLER_DATABASE_NAME, "", bucketName,
                         clientConfig);
            return false;
        }
```

- For API details, see GetCrawler in *AWS SDK for C++ API Reference*.

Java

SDK for Java 2.x

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
    /**
     * Retrieves a specific crawler from the AWS Glue service and waits for it to
be in the "READY" state.
     *
     * @param glueClient  the AWS Glue client used to interact with the Glue
service
     * @param crawlerName the name of the crawler to be retrieved
     */
    public static void getSpecificCrawler(GlueClient glueClient, String
crawlerName) throws InterruptedException {
        try {
            GetCrawlerRequest crawlerRequest = GetCrawlerRequest.builder()
                .name(crawlerName)
                .build();

            boolean ready = false;
            while (!ready) {
                GetCrawlerResponse response =
glueClient.getCrawler(crawlerRequest);
                String status = response.crawler().stateAsString();
                if (status.compareTo("READY") == 0) {
                    ready = true;
                }
                Thread.sleep(3000);
            }

            System.out.println("The crawler is now ready");

        } catch (GlueException | InterruptedException e) {
            throw e;
        }
    }
```

- For API details, see GetCrawler in *AWS SDK for Java 2.x API Reference*.

JavaScript

SDK for JavaScript (v3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
const getCrawler = (name) => {
  const client = new GlueClient({});

  const command = new GetCrawlerCommand({
    Name: name,
  });

  return client.send(command);
};
```

- For API details, see GetCrawler in *AWS SDK for JavaScript API Reference*.

Kotlin

SDK for Kotlin

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
suspend fun getSpecificCrawler(crawlerName: String?) {
    val request =
        GetCrawlerRequest {
            name = crawlerName
        }
    GlueClient.fromEnvironment { region = "us-east-1" }.use { glueClient ->
        val response = glueClient.getCrawler(request)
        val role = response.crawler?.role
        println("The role associated with this crawler is $role")
    }
}
```

- For API details, see GetCrawler in *AWS SDK for Kotlin API reference*.

PHP

SDK for PHP

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        echo "Waiting for crawler";
        do {
            $crawler = $glueService->getCrawler($crawlerName);
            echo ".";
            sleep(10);
        } while ($crawler['Crawler']['State'] != "READY");
        echo "\n";

    public function getCrawler($crawlerName)
    {
        return $this->customWaiter(function () use ($crawlerName) {
            return $this->glueClient->getCrawler([
                'Name' => $crawlerName,
            ]);
        });
    }
```

- For API details, see GetCrawler in *AWS SDK for PHP API Reference*.

Python

SDK for Python (Boto3)

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""

    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 Glue client.
        """
        self.glue_client = glue_client


    def get_crawler(self, name):
        """
        Gets information about a crawler.

        :param name: The name of the crawler to look up.
        :return: Data about the crawler.
        """
        crawler = None
        try:
            response = self.glue_client.get_crawler(Name=name)
            crawler = response["Crawler"]
        except ClientError as err:
            if err.response["Error"]["Code"] == "EntityNotFoundException":
                logger.info("Crawler %s doesn't exist.", name)
            else:
                logger.error(
                    "Couldn't get crawler %s. Here's why: %s: %s",
                    name,
                    err.response["Error"]["Code"],
                    err.response["Error"]["Message"],
                )
                raise
        return crawler
```

- For API details, see GetCrawler in *AWS SDK for Python (Boto3) API Reference*.

Ruby

SDK for Ruby

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
# The `GlueWrapper` class serves as a wrapper around the AWS Glue API, providing
 a simplified interface for common operations.
# It encapsulates the functionality of the AWS SDK for Glue and provides methods
 for interacting with Glue crawlers, databases, tables, jobs, and S3 resources.
# The class initializes with a Glue client and a logger, allowing it to make API
 calls and log any errors or informational messages.
class GlueWrapper
  def initialize(glue_client, logger)
    @glue_client = glue_client
    @logger = logger
  end

  # Retrieves information about a specific crawler.
  #
  # @param name [String] The name of the crawler to retrieve information about.
  # @return [Aws::Glue::Types::Crawler, nil] The crawler object if found, or nil
 if not found.
  def get_crawler(name)
    @glue_client.get_crawler(name: name)
  rescue Aws::Glue::Errors::EntityNotFoundException
    @logger.info("Crawler #{name} doesn't exist.")
    false
  rescue Aws::Glue::Errors::GlueException => e
    @logger.error("Glue could not get crawler #{name}: \n#{e.message}")
    raise
  end
```

- For API details, see GetCrawler in *AWS SDK for Ruby API Reference*.

Rust

SDK for Rust

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
            let tmp_crawler = glue
                .get_crawler()
                .name(self.crawler())
                .send()
                .await
                .map_err(GlueMvpError::from_glue_sdk)?;
```

- For API details, see GetCrawler in *AWS SDK for Rust API reference*.

Swift

SDK for Swift

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
import AWSClientRuntime
import AWSGlue

    /// Get the state of the specified AWS Glue crawler.
    ///
    /// - Parameters:
    ///   - glueClient: The AWS Glue client to use.
    ///   - name: The name of the crawler whose state should be returned.
    ///
    /// - Returns: A `GlueClientTypes.CrawlerState` value describing the
    ///   state of the crawler.
    func getCrawlerState(glueClient: GlueClient, name: String) async ->
 GlueClientTypes.CrawlerState {
        do {
            let output = try await glueClient.getCrawler(
                input: GetCrawlerInput(name: name)
            )

            // If the crawler or its state is `nil`, report that the crawler
            // is stopping. This may not be what you want for your
            // application but it works for this one!

            guard let crawler = output.crawler else {
                return GlueClientTypes.CrawlerState.stopping
            }
            guard let state = crawler.state else {
                return GlueClientTypes.CrawlerState.stopping
            }
            return state
        } catch {
            return GlueClientTypes.CrawlerState.stopping
        }
    }
```

- For API details, see GetCrawler in *AWS SDK for Swift API reference*.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `GetDatabase` with an AWS SDK

The following code examples show how to use `GetDatabase`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Learn the basics

.NET

SDK for .NET

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
    /// <summary>
    /// Get information about an AWS Glue database.
    /// </summary>
    /// <param name="dbName">The name of the database.</param>
    /// <returns>A Database object containing information about the database.</
returns>
    public async Task<Database> GetDatabaseAsync(string dbName)
    {
        var databasesRequest = new GetDatabaseRequest
        {
            Name = dbName,
        };

        var response = await _amazonGlue.GetDatabaseAsync(databasesRequest);
        return response.Database;
    }
```

- For API details, see GetDatabase in *AWS SDK for .NET API Reference*.

C++

SDK for C++

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the [AWS Code Examples Repository](https://github.com/awsdocs/aws-doc-sdk-examples/tree/main/cpp/example_code/glue#code-examples).

```
       Aws::Client::ClientConfiguration clientConfig;
       // Optional: Set to the AWS Region in which the bucket was created
(overrides config file).
       // clientConfig.region = "us-east-1";

   Aws::Glue::GlueClient client(clientConfig);

       Aws::Glue::Model::GetDatabaseRequest request;
       request.SetName(CRAWLER_DATABASE_NAME);

       Aws::Glue::Model::GetDatabaseOutcome outcome =
client.GetDatabase(request);

       if (outcome.IsSuccess()) {
           const Aws::Glue::Model::Database &database =
outcome.GetResult().GetDatabase();

           std::cout << "Successfully retrieve the database\n" <<
                     database.Jsonize().View().WriteReadable() << "'." <<
std::endl;
       }
       else {
           std::cerr << "Error getting the database.  "
                     << outcome.GetError().GetMessage() << std::endl;
           deleteAssets(CRAWLER_NAME, CRAWLER_DATABASE_NAME, "", bucketName,
                        clientConfig);
           return false;
       }
```

- For API details, see [GetDatabase](https://docs.aws.amazon.com/goto/SdkForCpp/glue-2017-03-31/GetDatabase) in *AWS SDK for C++ API Reference*.

Java

SDK for Java 2.x

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
    /**
     * Retrieves the specific database from the AWS Glue service.
     *
     * @param glueClient   an instance of the AWS Glue client used to interact
 with the service
     * @param databaseName the name of the database to retrieve
     * @throws GlueException if there is an error retrieving the database from
 the AWS Glue service
     */
    public static void getSpecificDatabase(GlueClient glueClient, String
 databaseName) {
        try {
            GetDatabaseRequest databasesRequest = GetDatabaseRequest.builder()
                .name(databaseName)
                .build();

            GetDatabaseResponse response =
 glueClient.getDatabase(databasesRequest);
            Instant createDate = response.database().createTime();

            // Convert the Instant to readable date.
            DateTimeFormatter formatter =
 DateTimeFormatter.ofLocalizedDateTime(FormatStyle.SHORT)
                .withLocale(Locale.US)
                .withZone(ZoneId.systemDefault());

            formatter.format(createDate);
            System.out.println("The create date of the database is " +
 createDate);

        } catch (GlueException e) {
            throw e;
        }
    }
```

- For API details, see GetDatabase in *AWS SDK for Java 2.x API Reference*.

JavaScript

SDK for JavaScript (v3)

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
const getDatabase = (name) => {
  const client = new GlueClient({});

  const command = new GetDatabaseCommand({
    Name: name,
  });

  return client.send(command);
};
```

- For API details, see GetDatabase in *AWS SDK for JavaScript API Reference*.

Kotlin

SDK for Kotlin

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
suspend fun getSpecificDatabase(databaseName: String?) {
    val request =
        GetDatabaseRequest {
            name = databaseName
        }

    GlueClient.fromEnvironment { region = "us-east-1" }.use { glueClient ->
        val response = glueClient.getDatabase(request)
        val dbDesc = response.database?.description
        println("The database description is $dbDesc")
    }
}
```

- For API details, see GetDatabase in *AWS SDK for Kotlin API reference*.

PHP

SDK for PHP

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        $databaseName = "doc-example-database-$uniqid";

        $database = $glueService->getDatabase($databaseName);
        echo "Found a database named " . $database['Database']['Name'] . "\n";

    public function getDatabase(string $databaseName): Result
    {
        return $this->customWaiter(function () use ($databaseName) {
            return $this->glueClient->getDatabase([
                'Name' => $databaseName,
            ]);
        });
    }
```

- For API details, see GetDatabase in *AWS SDK for PHP API Reference*.

Python

SDK for Python (Boto3)

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""

    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 Glue client.
        """
        self.glue_client = glue_client


    def get_database(self, name):
        """
        Gets information about a database in your Data Catalog.

        :param name: The name of the database to look up.
        :return: Information about the database.
        """
        try:
            response = self.glue_client.get_database(Name=name)
        except ClientError as err:
            logger.error(
                "Couldn't get database %s. Here's why: %s: %s",
                name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise
        else:
            return response["Database"]
```

- For API details, see GetDatabase in *AWS SDK for Python (Boto3) API Reference*.

Ruby

SDK for Ruby

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
# The `GlueWrapper` class serves as a wrapper around the AWS Glue API, providing
 a simplified interface for common operations.
# It encapsulates the functionality of the AWS SDK for Glue and provides methods
 for interacting with Glue crawlers, databases, tables, jobs, and S3 resources.
# The class initializes with a Glue client and a logger, allowing it to make API
 calls and log any errors or informational messages.
class GlueWrapper
  def initialize(glue_client, logger)
    @glue_client = glue_client
    @logger = logger
  end

  # Retrieves information about a specific database.
  #
  # @param name [String] The name of the database to retrieve information about.
  # @return [Aws::Glue::Types::Database, nil] The database object if found, or
 nil if not found.
  def get_database(name)
    response = @glue_client.get_database(name: name)
    response.database
  rescue Aws::Glue::Errors::GlueException => e
    @logger.error("Glue could not get database #{name}: \n#{e.message}")
    raise
  end
```

- For API details, see GetDatabase in *AWS SDK for Ruby API Reference*.

Rust

SDK for Rust

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        let database = glue
            .get_database()
            .name(self.database())
            .send()
            .await
            .map_err(GlueMvpError::from_glue_sdk)?
            .to_owned();
        let database = database
            .database()
            .ok_or_else(|| GlueMvpError::Unknown("Could not find 
database".into()))?;
```

- For API details, see GetDatabase in *AWS SDK for Rust API reference*.

Swift

SDK for Swift

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
import AWSClientRuntime
import AWSGlue
```

```
    /// Get the AWS Glue database with the specified name.
    ///
    /// - Parameters:
    ///   - glueClient: The AWS Glue client to use.
    ///   - name: The name of the database to return.
    ///
    /// - Returns: The `GlueClientTypes.Database` object describing the
    ///   specified database, or `nil` if an error occurs or the database
    ///   isn't found.
    func getDatabase(glueClient: GlueClient, name: String) async -> 
GlueClientTypes.Database? {
        do {
            let output = try await glueClient.getDatabase(
                input: GetDatabaseInput(name: name)
            )

            return output.database
        } catch {
            return nil
        }
    }
```

- For API details, see GetDatabase in *AWS SDK for Swift API reference*.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `GetDatabases` with an AWS SDK or CLI

The following code examples show how to use `GetDatabases`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Learn the basics

CLI

AWS CLI

To list the definitions of some or all of the databases in the AWS Glue Data Catalog

The following `get-databases` example returns information about the databases in the Data Catalog.

```
aws glue get-databases
```

Output:

```
{
    "DatabaseList": [
        {
            "Name": "default",
            "Description": "Default Hive database",
            "LocationUri": "file:/spark-warehouse",
            "CreateTime": 1602084052.0,
            "CreateTableDefaultPermissions": [
                {
                    "Principal": {
                        "DataLakePrincipalIdentifier": "IAM_ALLOWED_PRINCIPALS"
                    },
                    "Permissions": [
                        "ALL"
                    ]
                }
            ],
            "CatalogId": "111122223333"
        },
        {
            "Name": "flights-db",
            "CreateTime": 1587072847.0,
            "CreateTableDefaultPermissions": [
                {
                    "Principal": {
                        "DataLakePrincipalIdentifier": "IAM_ALLOWED_PRINCIPALS"
                    },
                    "Permissions": [
                        "ALL"
                    ]
                }
            ],
            "CatalogId": "111122223333"
        },
        {
            "Name": "legislators",
            "CreateTime": 1601415625.0,
            "CreateTableDefaultPermissions": [
                {
                    "Principal": {
                        "DataLakePrincipalIdentifier": "IAM_ALLOWED_PRINCIPALS"
                    },
                    "Permissions": [
                        "ALL"
                    ]
                }
            ],
            "CatalogId": "111122223333"
        },
        {
            "Name": "tempdb",
            "CreateTime": 1601498566.0,
            "CreateTableDefaultPermissions": [
                {
                    "Principal": {
                        "DataLakePrincipalIdentifier": "IAM_ALLOWED_PRINCIPALS"
                    },
                    "Permissions": [
                        "ALL"
                    ]
                }
            ],
            "CatalogId": "111122223333"
        }
    ]
}
```

For more information, see Defining a Database in Your Data Catalog in the *AWS Glue Developer Guide*.

- For API details, see GetDatabases in *AWS CLI Command Reference*.

JavaScript

SDK for JavaScript (v3)

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
const getDatabases = () => {
  const client = new GlueClient({});

  const command = new GetDatabasesCommand({});

  return client.send(command);
};
```

- For API details, see GetDatabases in *AWS SDK for JavaScript API Reference*.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `GetJob` with an AWS SDK or CLI

The following code examples show how to use `GetJob`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Learn the basics

CLI

AWS CLI

To retrieve information about a job

The following `get-job` example retrieves information about a job.

```
aws glue get-job \
    --job-name my-testing-job
```

Output:

```
{
    "Job": {
        "Name": "my-testing-job",
        "Role": "Glue_DefaultRole",
        "CreatedOn": 1602805698.167,
        "LastModifiedOn": 1602805698.167,
        "ExecutionProperty": {
            "MaxConcurrentRuns": 1
        },
        "Command": {
            "Name": "gluestreaming",
            "ScriptLocation": "s3://janetst-bucket-01/Scripts/test_script.scala",
            "PythonVersion": "2"
        },
        "DefaultArguments": {
            "--class": "GlueApp",
            "--job-language": "scala"
        },
        "MaxRetries": 0,
        "AllocatedCapacity": 10,
        "MaxCapacity": 10.0,
        "GlueVersion": "1.0"
    }
}
```

For more information, see Jobs in the *AWS Glue Developer Guide*.

- For API details, see GetJob in *AWS CLI Command Reference*.

JavaScript

SDK for JavaScript (v3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
const getJob = (jobName) => {
  const client = new GlueClient({});

  const command = new GetJobCommand({
    JobName: jobName,
  });

  return client.send(command);
};
```

- For API details, see GetJob in *AWS SDK for JavaScript API Reference*.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `GetJobRun` with an AWS SDK or CLI

The following code examples show how to use `GetJobRun`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Learn the basics

.NET

SDK for .NET

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
    /// <summary>
    /// Get information about a specific AWS Glue job run.
    /// </summary>
    /// <param name="jobName">The name of the job.</param>
    /// <param name="jobRunId">The Id of the job run.</param>
    /// <returns>A JobRun object with information about the job run.</returns>
    public async Task<JobRun> GetJobRunAsync(string jobName, string jobRunId)
    {
        var response = await _amazonGlue.GetJobRunAsync(new GetJobRunRequest
 { JobName = jobName, RunId = jobRunId });
        return response.JobRun;
    }
```

- For API details, see GetJobRun in *AWS SDK for .NET API Reference*.

C++

SDK for C++

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region in which the bucket was created
 (overrides config file).
```

```
        // clientConfig.region = "us-east-1";

    Aws::Glue::GlueClient client(clientConfig);

        Aws::Glue::Model::GetJobRunRequest jobRunRequest;
        jobRunRequest.SetJobName(jobName);
        jobRunRequest.SetRunId(jobRunID);

        Aws::Glue::Model::GetJobRunOutcome jobRunOutcome = client.GetJobRun(
                jobRunRequest);

        if (jobRunOutcome.IsSuccess()) {
            std::cout << "Displaying the job run JSON description." << std::endl;
            std::cout
                    <<
 jobRunOutcome.GetResult().GetJobRun().Jsonize().View().WriteReadable()
                    << std::endl;
        }
        else {
            std::cerr << "Error get a job run. "
                      << jobRunOutcome.GetError().GetMessage()
                      << std::endl;
        }
```

- For API details, see GetJobRun in *AWS SDK for C++ API Reference*.

CLI

AWS CLI

To get information about a job run

The following `get-job-run` example retrieves information about a job run.

```
aws glue get-job-run \
    --job-name "Combine legistators data" \
    --run-id jr_012e176506505074d94d761755e5c62538ee1aad6f17d39f527e9140cf0c9a5e
```

Output:

```
{
    "JobRun": {
        "Id": 
 "jr_012e176506505074d94d761755e5c62538ee1aad6f17d39f527e9140cf0c9a5e",
        "Attempt": 0,
        "JobName": "Combine legistators data",
        "StartedOn": 1602873931.255,
        "LastModifiedOn": 1602874075.985,
        "CompletedOn": 1602874075.985,
        "JobRunState": "SUCCEEDED",
        "Arguments": {
            "--enable-continuous-cloudwatch-log": "true",
            "--enable-metrics": "",
            "--enable-spark-ui": "true",
            "--job-bookmark-option": "job-bookmark-enable",
            "--spark-event-logs-path": "s3://aws-glue-assets-111122223333-us-
east-1/sparkHistoryLogs/"
        },
        "PredecessorRuns": [],
        "AllocatedCapacity": 10,
        "ExecutionTime": 117,
        "Timeout": 2880,
        "MaxCapacity": 10.0,
        "WorkerType": "G.1X",
        "NumberOfWorkers": 10,
        "LogGroupName": "/aws-glue/jobs",
        "GlueVersion": "2.0"
    }
}
```

For more information, see Job Runs in the *AWS Glue Developer Guide*.

- For API details, see GetJobRun in *AWS CLI Command Reference*.

JavaScript

SDK for JavaScript (v3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
const getJobRun = (jobName, jobRunId) => {
  const client = new GlueClient({});
  const command = new GetJobRunCommand({
    JobName: jobName,
    RunId: jobRunId,
  });

  return client.send(command);
};
```

- For API details, see GetJobRun in *AWS SDK for JavaScript API Reference*.

PHP

SDK for PHP

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        $jobName = 'test-job-' . $uniqid;

        $outputBucketUrl = "s3://$bucketName";
        $runId = $glueService->startJobRun($jobName, $databaseName, $tables,
 $outputBucketUrl)['JobRunId'];

        echo "waiting for job";
        do {
            $jobRun = $glueService->getJobRun($jobName, $runId);
            echo ".";
            sleep(10);
        } while (!array_intersect([$jobRun['JobRun']['JobRunState']],
 ['SUCCEEDED', 'STOPPED', 'FAILED', 'TIMEOUT']));
        echo "\n";

    public function getJobRun($jobName, $runId, $predecessorsIncluded = false):
 Result
    {
        return $this->glueClient->getJobRun([
            'JobName' => $jobName,
            'RunId' => $runId,
            'PredecessorsIncluded' => $predecessorsIncluded,
        ]);
    }
```

- For API details, see GetJobRun in *AWS SDK for PHP API Reference*.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""

    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 Glue client.
        """
        self.glue_client = glue_client


    def get_job_run(self, name, run_id):
        """
        Gets information about a single job run.

        :param name: The name of the job definition for the run.
        :param run_id: The ID of the run.
        :return: Information about the run.
        """
        try:
            response = self.glue_client.get_job_run(JobName=name, RunId=run_id)
        except ClientError as err:
            logger.error(
                "Couldn't get job run %s/%s. Here's why: %s: %s",
                name,
                run_id,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise
        else:
            return response["JobRun"]
```

- For API details, see GetJobRun in *AWS SDK for Python (Boto3) API Reference*.

Ruby

SDK for Ruby

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
# The `GlueWrapper` class serves as a wrapper around the AWS Glue API, providing
 a simplified interface for common operations.
# It encapsulates the functionality of the AWS SDK for Glue and provides methods
 for interacting with Glue crawlers, databases, tables, jobs, and S3 resources.
# The class initializes with a Glue client and a logger, allowing it to make API
 calls and log any errors or informational messages.
class GlueWrapper
  def initialize(glue_client, logger)
    @glue_client = glue_client
    @logger = logger
  end

  # Retrieves data for a specific job run.
  #
  # @param job_name [String] The name of the job run to retrieve data for.
  # @return [Glue::Types::GetJobRunResponse]
  def get_job_run(job_name, run_id)
```

```
   @glue_client.get_job_run(job_name: job_name, run_id: run_id)
 rescue Aws::Glue::Errors::GlueException => e
   @logger.error("Glue could not get job runs: \n#{e.message}")
 end
```

- For API details, see GetJobRun in *AWS SDK for Ruby API Reference*.

Rust

SDK for Rust

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
       let get_job_run = || async {
           Ok::<JobRun, GlueMvpError>(
               glue.get_job_run()
                   .job_name(self.job())
                   .run_id(job_run_id.to_string())
                   .send()
                   .await
                   .map_err(GlueMvpError::from_glue_sdk)?
                   .job_run()
                   .ok_or_else(|| GlueMvpError::Unknown("Failed to get
job_run".into()))?
                   .to_owned(),
           )
       };

       let mut job_run = get_job_run().await?;
       let mut state =
job_run.job_run_state().unwrap_or(&unknown_state).to_owned();

       while matches!(
           state,
           JobRunState::Starting | JobRunState::Stopping | JobRunState::Running
       ) {
           info!(?state, "Waiting for job to finish");
```

```
            tokio::time::sleep(self.wait_delay).await;

            job_run = get_job_run().await?;
            state = job_run.job_run_state().unwrap_or(&unknown_state).to_owned();
        }
```

- For API details, see GetJobRun in *AWS SDK for Rust API reference*.

Swift

SDK for Swift

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
import AWSClientRuntime
import AWSGlue

    /// Get information about a specific AWS Glue job run.
    ///
    /// - Parameters:
    ///   - glueClient: The AWS Glue client to use.
    ///   - jobName: The name of the job to return job run data for.
    ///   - id: The run ID of the specific job run to return.
    ///
    /// - Returns: A `GlueClientTypes.JobRun` object describing the state of
    ///   the job run, or `nil` if an error occurs.
    func getJobRun(glueClient: GlueClient, name jobName: String, id: String)
 async -> GlueClientTypes.JobRun? {
        do {
            let output = try await glueClient.getJobRun(
                input: GetJobRunInput(
                    jobName: jobName,
                    runId: id
                )
            )

            return output.jobRun
        } catch {
            return nil
        }
    }
```

- For API details, see GetJobRun in *AWS SDK for Swift API reference*.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `GetJobRuns` with an AWS SDK or CLI

The following code examples show how to use `GetJobRuns`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Learn the basics

.NET

SDK for .NET

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
    /// <summary>
    /// Get information about all AWS Glue runs of a specific job.
    /// </summary>
    /// <param name="jobName">The name of the job.</param>
    /// <returns>A list of JobRun objects.</returns>
    public async Task<List<JobRun>> GetJobRunsAsync(string jobName)
    {
        var jobRuns = new List<JobRun>();
        var request = new GetJobRunsRequest
        {
            JobName = jobName,
        };

        // No need to loop to get all the log groups--the SDK does it for us
behind the scenes
        var paginatorForJobRuns =
            _amazonGlue.Paginators.GetJobRuns(request);

        await foreach (var response in paginatorForJobRuns.Responses)
        {
            response.JobRuns.ForEach(jobRun =>
            {
                jobRuns.Add(jobRun);
            });
        }

        return jobRuns;
    }
```

- For API details, see GetJobRuns in *AWS SDK for .NET API Reference*.

C++

SDK for C++

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region in which the bucket was created
(overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::Glue::GlueClient client(clientConfig);
```

```
        Aws::Glue::Model::GetJobRunsRequest getJobRunsRequest;
        getJobRunsRequest.SetJobName(jobName);

        Aws::String nextToken; // Used for pagination.
        std::vector<Aws::Glue::Model::JobRun> allJobRuns;
        do {
            if (!nextToken.empty()) {
                getJobRunsRequest.SetNextToken(nextToken);
            }
            Aws::Glue::Model::GetJobRunsOutcome jobRunsOutcome =
client.GetJobRuns(
                    getJobRunsRequest);

            if (jobRunsOutcome.IsSuccess()) {
                const std::vector<Aws::Glue::Model::JobRun> &jobRuns =
jobRunsOutcome.GetResult().GetJobRuns();
                allJobRuns.insert(allJobRuns.end(), jobRuns.begin(),
jobRuns.end());

                nextToken = jobRunsOutcome.GetResult().GetNextToken();
            }
            else {
                std::cerr << "Error getting job runs. "
                          << jobRunsOutcome.GetError().GetMessage()
                          << std::endl;
                break;
            }
        } while (!nextToken.empty());
```

- For API details, see GetJobRuns in *AWS SDK for C++ API Reference*.

CLI

AWS CLI

To get information about all job runs for a job

The following `get-job-runs` example retrieves information about job runs for a job.

```
aws glue get-job-runs \
    --job-name "my-testing-job"
```

Output:

```
{
    "JobRuns": [
        {
            "Id":
 "jr_012e176506505074d94d761755e5c62538ee1aad6f17d39f527e9140cf0c9a5e",
            "Attempt": 0,
            "JobName": "my-testing-job",
            "StartedOn": 1602873931.255,
            "LastModifiedOn": 1602874075.985,
            "CompletedOn": 1602874075.985,
            "JobRunState": "SUCCEEDED",
            "Arguments": {
                "--enable-continuous-cloudwatch-log": "true",
                "--enable-metrics": "",
                "--enable-spark-ui": "true",
                "--job-bookmark-option": "job-bookmark-enable",
                "--spark-event-logs-path": "s3://aws-glue-assets-111122223333-us-
east-1/sparkHistoryLogs/"
            },
            "PredecessorRuns": [],
            "AllocatedCapacity": 10,
            "ExecutionTime": 117,
            "Timeout": 2880,
            "MaxCapacity": 10.0,
            "WorkerType": "G.1X",
            "NumberOfWorkers": 10,
            "LogGroupName": "/aws-glue/jobs",
            "GlueVersion": "2.0"
        },
        {
            "Id":
 "jr_03cc19ddab11c4e244d3f735567de74ff93b0b3ef468a713ffe73e53d1aec08f_attempt_2",
            "Attempt": 2,
            "PreviousRunId":
 "jr_03cc19ddab11c4e244d3f735567de74ff93b0b3ef468a713ffe73e53d1aec08f_attempt_1",
            "JobName": "my-testing-job",
            "StartedOn": 1602811168.496,
            "LastModifiedOn": 1602811282.39,
            "CompletedOn": 1602811282.39,
            "JobRunState": "FAILED",
            "ErrorMessage": "An error occurred while calling
 o122.pyWriteDynamicFrame.
```

```
                Access Denied (Service: Amazon S3; Status Code: 403; Error Code:
 AccessDenied;
                Request ID: 021AAB703DB20A2D;
                S3 Extended Request ID: teZk24Y09TkXzBvMPG502L5VJBhe9DJuWA9/
TXtuGOqfByajkfL/Tlqt5JBGdEGpigAqzdMDM/U=)",
            "PredecessorRuns": [],
            "AllocatedCapacity": 10,
            "ExecutionTime": 110,
            "Timeout": 2880,
            "MaxCapacity": 10.0,
            "WorkerType": "G.1X",
            "NumberOfWorkers": 10,
            "LogGroupName": "/aws-glue/jobs",
            "GlueVersion": "2.0"
        },
        {
            "Id":
 "jr_03cc19ddab11c4e244d3f735567de74ff93b0b3ef468a713ffe73e53d1aec08f_attempt_1",
            "Attempt": 1,
            "PreviousRunId":
 "jr_03cc19ddab11c4e244d3f735567de74ff93b0b3ef468a713ffe73e53d1aec08f",
            "JobName": "my-testing-job",
            "StartedOn": 1602811020.518,
            "LastModifiedOn": 1602811138.364,
            "CompletedOn": 1602811138.364,
            "JobRunState": "FAILED",
            "ErrorMessage": "An error occurred while calling
 o122.pyWriteDynamicFrame.
                 Access Denied (Service: Amazon S3; Status Code: 403; Error Code:
 AccessDenied;
                 Request ID: 2671D37856AE7ABB;
                 S3 Extended Request ID: RLJCJw20brV
+PpC6GpORahyF2fp9flB5SSb2bTGPnUSPVizLXRl1PN3QZldb+v1o9qRVktNYbW8=)",
            "PredecessorRuns": [],
            "AllocatedCapacity": 10,
            "ExecutionTime": 113,
            "Timeout": 2880,
            "MaxCapacity": 10.0,
            "WorkerType": "G.1X",
            "NumberOfWorkers": 10,
            "LogGroupName": "/aws-glue/jobs",
            "GlueVersion": "2.0"
        }
    ]
```

```
}
```

For more information, see [Job Runs](#) in the *AWS Glue Developer Guide*.

- For API details, see [GetJobRuns](#) in *AWS CLI Command Reference*.

Java

SDK for Java 2.x

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the [AWS Code Examples Repository](#).

```
    /**
     * Retrieves the job runs for a given Glue job and prints the status of the
 job runs.
     *
     * @param glueClient the Glue client used to make API calls
     * @param jobName    the name of the Glue job to retrieve the job runs for
     */
    public static void getJobRuns(GlueClient glueClient, String jobName) {
        try {
            GetJobRunsRequest runsRequest = GetJobRunsRequest.builder()
                .jobName(jobName)
                .maxResults(20)
                .build();

            boolean jobDone = false;
            while (!jobDone) {
                GetJobRunsResponse response = glueClient.getJobRuns(runsRequest);
                List<JobRun> jobRuns = response.jobRuns();
                for (JobRun jobRun : jobRuns) {
                    String jobState = jobRun.jobRunState().name();
                    if (jobState.compareTo("SUCCEEDED") == 0) {
                        System.out.println(jobName + " has succeeded");
                        jobDone = true;

                    } else if (jobState.compareTo("STOPPED") == 0) {
                        System.out.println("Job run has stopped");
                        jobDone = true;

                    } else if (jobState.compareTo("FAILED") == 0) {
                        System.out.println("Job run has failed");
                        jobDone = true;

                    } else if (jobState.compareTo("TIMEOUT") == 0) {
                        System.out.println("Job run has timed out");
                        jobDone = true;

                    } else {
                        System.out.println("*** Job run state is " +
jobRun.jobRunState().name());
                        System.out.println("Job run Id is " + jobRun.id());
                        System.out.println("The Glue version is " +
jobRun.glueVersion());
                    }
                    TimeUnit.SECONDS.sleep(5);
                }
            }

        } catch (GlueException e) {
            throw e;
        } catch (InterruptedException e) {
            throw new RuntimeException(e);
        }
    }
```

- For API details, see GetJobRuns in *AWS SDK for Java 2.x API Reference*.

JavaScript

SDK for JavaScript (v3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
const getJobRuns = (jobName) => {
  const client = new GlueClient({});
  const command = new GetJobRunsCommand({
    JobName: jobName,
  });

  return client.send(command);
};
```

- For API details, see GetJobRuns in *AWS SDK for JavaScript API Reference*.

PHP

SDK for PHP

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        $jobName = 'test-job-' . $uniqid;

        $jobRuns = $glueService->getJobRuns($jobName);

    public function getJobRuns($jobName, $maxResults = 0, $nextToken = ''):
 Result
    {
        $arguments = ['JobName' => $jobName];
        if ($maxResults) {
            $arguments['MaxResults'] = $maxResults;
        }
        if ($nextToken) {
            $arguments['NextToken'] = $nextToken;
        }
        return $this->glueClient->getJobRuns($arguments);
    }
```

- For API details, see GetJobRuns in *AWS SDK for PHP API Reference*.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""

    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 Glue client.
        """
        self.glue_client = glue_client


    def get_job_runs(self, job_name):
        """
        Gets information about runs that have been performed for a specific job
        definition.

        :param job_name: The name of the job definition to look up.
        :return: The list of job runs.
        """
        try:
            response = self.glue_client.get_job_runs(JobName=job_name)
        except ClientError as err:
            logger.error(
                "Couldn't get job runs for %s. Here's why: %s: %s",
                job_name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise
        else:
            return response["JobRuns"]
```

- For API details, see GetJobRuns in *AWS SDK for Python (Boto3) API Reference*.

Ruby

SDK for Ruby

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
# The `GlueWrapper` class serves as a wrapper around the AWS Glue API, providing
 a simplified interface for common operations.
# It encapsulates the functionality of the AWS SDK for Glue and provides methods
 for interacting with Glue crawlers, databases, tables, jobs, and S3 resources.
# The class initializes with a Glue client and a logger, allowing it to make API
 calls and log any errors or informational messages.
class GlueWrapper
  def initialize(glue_client, logger)
    @glue_client = glue_client
    @logger = logger
  end

  # Retrieves a list of job runs for the specified job.
  #
  # @param job_name [String] The name of the job to retrieve job runs for.
  # @return [Array<Aws::Glue::Types::JobRun>]
  def get_job_runs(job_name)
    response = @glue_client.get_job_runs(job_name: job_name)
    response.job_runs
  rescue Aws::Glue::Errors::GlueException => e
    @logger.error("Glue could not get job runs: \n#{e.message}")
  end
```

- For API details, see GetJobRuns in *AWS SDK for Ruby API Reference*.

Swift

SDK for Swift

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
import AWSClientRuntime
import AWSGlue

    /// Return a list of the job runs for the specified job.
    ///
    /// - Parameters:
    ///   - glueClient: The AWS Glue client to use.
    ///   - jobName: The name of the job for which to return its job runs.
    ///   - maxResults: The maximum number of job runs to return (default:
    ///     1000).
    ///
    /// - Returns: An array of `GlueClientTypes.JobRun` objects describing
    ///   each job run.
    func getJobRuns(glueClient: GlueClient, name jobName: String, maxResults:
 Int? = nil) async -> [GlueClientTypes.JobRun] {
        do {
            let output = try await glueClient.getJobRuns(
                input: GetJobRunsInput(
                    jobName: jobName,
                    maxResults: maxResults
                )
            )

            guard let jobRuns = output.jobRuns else {
                print("*** No job runs found.")
                return []
            }

            return jobRuns
        } catch is EntityNotFoundException {
            print("*** The specified job name, \(jobName), doesn't exist.")
            return []
        } catch {
            print("*** Unexpected error getting job runs:")
            dump(error)
            return []
        }
    }
```

- For API details, see GetJobRuns in *AWS SDK for Swift API reference*.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `GetTables` with an AWS SDK or CLI

The following code examples show how to use `GetTables`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Learn the basics

.NET

SDK for .NET

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
    /// <summary>
    /// Get a list of tables for an AWS Glue database.
    /// </summary>
    /// <param name="dbName">The name of the database.</param>
    /// <returns>A list of Table objects.</returns>
    public async Task<List<Table>> GetTablesAsync(string dbName)
    {
        var request = new GetTablesRequest { DatabaseName = dbName };
        var tables = new List<Table>();

        // Get a paginator for listing the tables.
        var tablePaginator = _amazonGlue.Paginators.GetTables(request);

        await foreach (var response in tablePaginator.Responses)
        {
            tables.AddRange(response.TableList);
        }

        return tables;
    }
```

- For API details, see GetTables in *AWS SDK for .NET API Reference*.

C++

SDK for C++

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region in which the bucket was created
(overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::Glue::GlueClient client(clientConfig);

        Aws::Glue::Model::GetTablesRequest request;
        request.SetDatabaseName(CRAWLER_DATABASE_NAME);
        std::vector<Aws::Glue::Model::Table> all_tables;
        Aws::String nextToken; // Used for pagination.
        do {
            Aws::Glue::Model::GetTablesOutcome outcome =
client.GetTables(request);
```

```
           if (outcome.IsSuccess()) {
               const std::vector<Aws::Glue::Model::Table> &tables =
outcome.GetResult().GetTableList();
               all_tables.insert(all_tables.end(), tables.begin(),
tables.end());
               nextToken = outcome.GetResult().GetNextToken();
           }
           else {
               std::cerr << "Error getting the tables. "
                         << outcome.GetError().GetMessage()
                         << std::endl;
               deleteAssets(CRAWLER_NAME, CRAWLER_DATABASE_NAME, "", bucketName,
                            clientConfig);
               return false;
           }
       } while (!nextToken.empty());

       std::cout << "The database contains " << all_tables.size()
                 << (all_tables.size() == 1 ?
                     " table." : "tables.") << std::endl;
       std::cout << "Here is a list of the tables in the database.";
       for (size_t index = 0; index < all_tables.size(); ++index) {
           std::cout << "    " << index + 1 << ":  " <<
all_tables[index].GetName()
                     << std::endl;
       }

       if (!all_tables.empty()) {
           int tableIndex = askQuestionForIntRange(
                   "Enter an index to display the database detail ",
                   1, static_cast<int>(all_tables.size()));
           std::cout << all_tables[tableIndex -
1].Jsonize().View().WriteReadable()
                     << std::endl;

           tableName = all_tables[tableIndex - 1].GetName();
       }
```

- For API details, see GetTables in *AWS SDK for C++ API Reference*.

CLI

AWS CLI

To list the definitions of some or all of the tables in the specified database

The following `get-tables` example returns information about the tables in the specified database.

```
aws glue get-tables --database-name 'tempdb'
```

Output:

```
{
    "TableList": [
        {
            "Name": "my-s3-sink",
            "DatabaseName": "tempdb",
            "CreateTime": 1602730539.0,
            "UpdateTime": 1602730539.0,
            "Retention": 0,
            "StorageDescriptor": {
                "Columns": [
                    {
                        "Name": "sensorid",
                        "Type": "int"
                    },
                    {
                        "Name": "currenttemperature",
                        "Type": "int"
                    },
                    {
                        "Name": "status",
                        "Type": "string"
                    }
                ],
                "Location": "s3://janetst-bucket-01/test-s3-output/",
                "Compressed": false,
                "NumberOfBuckets": 0,
                "SerdeInfo": {
                    "SerializationLibrary": "org.openx.data.jsonserde.JsonSerDe"
                },
                "SortColumns": [],
                "StoredAsSubDirectories": false
            },
            "Parameters": {
                "classification": "json"
            },
            "CreatedBy": "arn:aws:iam::007436865787:user/JRSTERN",
            "IsRegisteredWithLakeFormation": false,
            "CatalogId": "007436865787"
        },
        {
            "Name": "s3-source",
            "DatabaseName": "tempdb",
            "CreateTime": 1602730658.0,
            "UpdateTime": 1602730658.0,
            "Retention": 0,
            "StorageDescriptor": {
                "Columns": [
                    {
                        "Name": "sensorid",
                        "Type": "int"
                    },
                    {
                        "Name": "currenttemperature",
                        "Type": "int"
                    },
                    {
                        "Name": "status",
                        "Type": "string"
                    }
                ],
                "Location": "s3://janetst-bucket-01/",
                "Compressed": false,
                "NumberOfBuckets": 0,
                "SortColumns": [],
                "StoredAsSubDirectories": false
            },
            "Parameters": {
                "classification": "json"
            },
            "CreatedBy": "arn:aws:iam::007436865787:user/JRSTERN",
            "IsRegisteredWithLakeFormation": false,
            "CatalogId": "007436865787"
        },
        {
            "Name": "test-kinesis-input",
            "DatabaseName": "tempdb",
            "CreateTime": 1601507001.0,
            "UpdateTime": 1601507001.0,
            "Retention": 0,
            "StorageDescriptor": {
                "Columns": [
                    {
                        "Name": "sensorid",
                        "Type": "int"
                    },
                    {
                        "Name": "currenttemperature",
                        "Type": "int"
                    },
                    {
                        "Name": "status",
                        "Type": "string"
                    }
                ],
                "Location": "my-testing-stream",
                "Compressed": false,
                "NumberOfBuckets": 0,
                "SerdeInfo": {
                    "SerializationLibrary": "org.openx.data.jsonserde.JsonSerDe"
                },
                "SortColumns": [],
                "Parameters": {
                    "kinesisUrl": "https://kinesis.us-east-1.amazonaws.com",
                    "streamName": "my-testing-stream",
                    "typeOfData": "kinesis"
                },
                "StoredAsSubDirectories": false
            },
            "Parameters": {
                "classification": "json"
            },
            "CreatedBy": "arn:aws:iam::007436865787:user/JRSTERN",
            "IsRegisteredWithLakeFormation": false,
            "CatalogId": "007436865787"
        }
    ]
}
```

For more information, see Defining Tables in the AWS Glue Data Catalog in the *AWS Glue Developer Guide*.

- For API details, see GetTables in *AWS CLI Command Reference*.

Java

SDK for Java 2.x

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
    /**
     * Retrieves the names of the tables in the specified Glue database.
     *
     * @param glueClient the Glue client to use for the operation
     * @param dbName     the name of the Glue database to retrieve the table
 names from
     * @return the name of the first table retrieved, or an empty string if no
 tables were found
     */
    public static String getGlueTables(GlueClient glueClient, String dbName) {
        String myTableName = "";
        try {
            GetTablesRequest tableRequest = GetTablesRequest.builder()
                .databaseName(dbName)
                .build();

            GetTablesResponse response = glueClient.getTables(tableRequest);
            List<Table> tables = response.tableList();
            if (tables.isEmpty()) {
                System.out.println("No tables were returned");
            } else {
                for (Table table : tables) {
                    myTableName = table.name();
                    System.out.println("Table name is: " + myTableName);
                }
            }
        } catch (GlueException e) {
            throw e;
        }
        return myTableName;
    }
```

- For API details, see GetTables in *AWS SDK for Java 2.x API Reference*.

JavaScript

SDK for JavaScript (v3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
const getTables = (databaseName) => {
  const client = new GlueClient({});

  const command = new GetTablesCommand({
    DatabaseName: databaseName,
  });

  return client.send(command);
};
```

- For API details, see GetTables in *AWS SDK for JavaScript API Reference*.

PHP

SDK for PHP

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        $databaseName = "doc-example-database-$uniqid";

        $tables = $glueService->getTables($databaseName);

    public function getTables($databaseName): Result
    {
        return $this->glueClient->getTables([
            'DatabaseName' => $databaseName,
        ]);
    }
```

- For API details, see GetTables in *AWS SDK for PHP API Reference*.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""

    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 Glue client.
        """
        self.glue_client = glue_client


    def get_tables(self, db_name):
        """
        Gets a list of tables in a Data Catalog database.

        :param db_name: The name of the database to query.
        :return: The list of tables in the database.
        """
        try:
            response = self.glue_client.get_tables(DatabaseName=db_name)
        except ClientError as err:
            logger.error(
                "Couldn't get tables %s. Here's why: %s: %s",
                db_name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise
        else:
            return response["TableList"]
```

- For API details, see GetTables in *AWS SDK for Python (Boto3) API Reference*.

Ruby

SDK for Ruby

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
# The `GlueWrapper` class serves as a wrapper around the AWS Glue API, providing
 a simplified interface for common operations.
```

```
# It encapsulates the functionality of the AWS SDK for Glue and provides methods
 for interacting with Glue crawlers, databases, tables, jobs, and S3 resources.
# The class initializes with a Glue client and a logger, allowing it to make API
 calls and log any errors or informational messages.
class GlueWrapper
  def initialize(glue_client, logger)
    @glue_client = glue_client
    @logger = logger
  end

  # Retrieves a list of tables in the specified database.
  #
  # @param db_name [String] The name of the database to retrieve tables from.
  # @return [Array<Aws::Glue::Types::Table>]
  def get_tables(db_name)
    response = @glue_client.get_tables(database_name: db_name)
    response.table_list
  rescue Aws::Glue::Errors::GlueException => e
    @logger.error("Glue could not get tables #{db_name}: \n#{e.message}")
    raise
  end
```

- For API details, see GetTables in *AWS SDK for Ruby API Reference*.

Rust

SDK for Rust

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        let tables = glue
            .get_tables()
            .database_name(self.database())
            .send()
            .await
            .map_err(GlueMvpError::from_glue_sdk)?;
        let tables = tables.table_list();
```

- For API details, see GetTables in *AWS SDK for Rust API reference*.

Swift

SDK for Swift

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
import AWSClientRuntime
import AWSGlue

    /// Returns a list of the tables in the specified database.
    ///
    /// - Parameters:
    ///   - glueClient: The AWS Glue client to use.
    ///   - databaseName: The name of the database whose tables are to be
    ///     returned.
    ///
    /// - Returns: An array of `GlueClientTypes.Table` objects, each
    ///   describing one table in the named database. An empty array indicates
    ///   that there are either no tables in the database, or an error
    ///   occurred before any tables could be found.
    func getTablesInDatabase(glueClient: GlueClient, databaseName: String) async
 -> [GlueClientTypes.Table] {
        var tables: [GlueClientTypes.Table] = []
        var nextToken: String?

        repeat {
            do {
                let output = try await glueClient.getTables(
                    input: GetTablesInput(
                        databaseName: databaseName,
                        nextToken: nextToken
                    )
                )
                guard let tableList = output.tableList else {
                    return tables
                }

                tables = tables + tableList
                nextToken = output.nextToken
            } catch {
                return tables
            }
        } while nextToken != nil

        return tables
    }
```

- For API details, see GetTables in *AWS SDK for Swift API reference*.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `ListJobs` with an AWS SDK

The following code examples show how to use `ListJobs`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Learn the basics

.NET

SDK for .NET

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
    /// <summary>
    /// List AWS Glue jobs using a paginator.
    /// </summary>
    /// <returns>A list of AWS Glue job names.</returns>
    public async Task<List<string>> ListJobsAsync()
    {
        var jobNames = new List<string>();

        var listJobsPaginator = _amazonGlue.Paginators.ListJobs(new
 ListJobsRequest { MaxResults = 10 });
        await foreach (var response in listJobsPaginator.Responses)
        {
            jobNames.AddRange(response.JobNames);
        }

        return jobNames;
    }
```

- For API details, see ListJobs in *AWS SDK for .NET API Reference*.

C++

SDK for C++

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region in which the bucket was created
 (overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::Glue::GlueClient client(clientConfig);

        Aws::Glue::Model::ListJobsRequest listJobsRequest;
        Aws::String nextToken;
        std::vector<Aws::String> allJobNames;

        do {
            if (!nextToken.empty()) {
                listJobsRequest.SetNextToken(nextToken);
            }
            Aws::Glue::Model::ListJobsOutcome listRunsOutcome = client.ListJobs(
                    listJobsRequest);

            if (listRunsOutcome.IsSuccess()) {
                const std::vector<Aws::String> &jobNames =
 listRunsOutcome.GetResult().GetJobNames();
                allJobNames.insert(allJobNames.end(), jobNames.begin(),
 jobNames.end());
                nextToken = listRunsOutcome.GetResult().GetNextToken();
            }
            else {
                std::cerr << "Error listing jobs. "
                          << listRunsOutcome.GetError().GetMessage()
                          << std::endl;
            }
        } while (!nextToken.empty());
```

- For API details, see ListJobs in *AWS SDK for C++ API Reference*.

JavaScript

SDK for JavaScript (v3)

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
const listJobs = () => {
  const client = new GlueClient({});

  const command = new ListJobsCommand({});
  return client.send(command);
};
```

- For API details, see ListJobs in *AWS SDK for JavaScript API Reference*.

PHP

SDK for PHP

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        $jobs = $glueService->listJobs();
        echo "Current jobs:\n";
        foreach ($jobs['JobNames'] as $jobsName) {
            echo "{$jobsName}\n";
        }

    public function listJobs($maxResults = null, $nextToken = null, $tags = []):
 Result
    {
        $arguments = [];
        if ($maxResults) {
            $arguments['MaxResults'] = $maxResults;
        }
        if ($nextToken) {
            $arguments['NextToken'] = $nextToken;
        }
        if (!empty($tags)) {
            $arguments['Tags'] = $tags;
        }
        return $this->glueClient->listJobs($arguments);
    }
```

- For API details, see ListJobs in *AWS SDK for PHP API Reference*.

Python

SDK for Python (Boto3)

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""

    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 Glue client.
        """
        self.glue_client = glue_client


    def list_jobs(self):
        """
        Lists the names of job definitions in your account.

        :return: The list of job definition names.
        """
        try:
            response = self.glue_client.list_jobs()
        except ClientError as err:
            logger.error(
                "Couldn't list jobs. Here's why: %s: %s",
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise
        else:
            return response["JobNames"]
```

- For API details, see ListJobs in *AWS SDK for Python (Boto3) API Reference*.

Ruby

SDK for Ruby

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
# The `GlueWrapper` class serves as a wrapper around the AWS Glue API, providing
 a simplified interface for common operations.
# It encapsulates the functionality of the AWS SDK for Glue and provides methods
 for interacting with Glue crawlers, databases, tables, jobs, and S3 resources.
# The class initializes with a Glue client and a logger, allowing it to make API
 calls and log any errors or informational messages.
class GlueWrapper
  def initialize(glue_client, logger)
    @glue_client = glue_client
    @logger = logger
  end

  # Retrieves a list of jobs in AWS Glue.
  #
  # @return [Aws::Glue::Types::ListJobsResponse]
  def list_jobs
    @glue_client.list_jobs
  rescue Aws::Glue::Errors::GlueException => e
    @logger.error("Glue could not list jobs: \n#{e.message}")
    raise
  end
```

- For API details, see ListJobs in *AWS SDK for Ruby API Reference*.

Rust

SDK for Rust

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        let mut list_jobs = glue.list_jobs().into_paginator().send();
        while let Some(list_jobs_output) = list_jobs.next().await {
            match list_jobs_output {
                Ok(list_jobs) => {
                    let names = list_jobs.job_names();
                    info!(?names, "Found these jobs")
                }
                Err(err) => return Err(GlueMvpError::from_glue_sdk(err)),
            }
        }
```

- For API details, see ListJobs in *AWS SDK for Rust API reference*.

Swift

SDK for Swift

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
import AWSClientRuntime
import AWSGlue

    /// Return a list of the AWS Glue jobs listed on the user's account.
    ///
    /// - Parameters:
```

```
    ///   - glueClient: The AWS Glue client to use.
    ///   - maxJobs: The maximum number of jobs to return (default: 100).
    ///
    /// - Returns: An array of strings listing the names of all available AWS
    ///   Glue jobs.
    func listJobs(glueClient: GlueClient, maxJobs: Int = 100) async -> [String] {
        var jobList: [String] = []
        var nextToken: String?

        repeat {
            do {
                let output = try await glueClient.listJobs(
                    input: ListJobsInput(
                        maxResults: maxJobs,
                        nextToken: nextToken
                    )
                )

                guard let jobs = output.jobNames else {
                    return jobList
                }

                jobList = jobList + jobs
                nextToken = output.nextToken
            } catch {
                return jobList
            }
        } while (nextToken != nil)

        return jobList
    }
```

- For API details, see ListJobs in *AWS SDK for Swift API reference*.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `StartCrawler` with an AWS SDK or CLI

The following code examples show how to use `StartCrawler`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Learn the basics

.NET

SDK for .NET

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
/// <summary>
/// Start an AWS Glue crawler.
/// </summary>
/// <param name="crawlerName">The name of the crawler.</param>
/// <returns>A Boolean value indicating the success of the action.</returns>
public async Task<bool> StartCrawlerAsync(string crawlerName)
{
    var crawlerRequest = new StartCrawlerRequest
    {
        Name = crawlerName,
    };

    var response = await _amazonGlue.StartCrawlerAsync(crawlerRequest);

    return response.HttpStatusCode == System.Net.HttpStatusCode.OK;
}
```

- For API details, see StartCrawler in *AWS SDK for .NET API Reference*.

C++

SDK for C++

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the [AWS Code Examples Repository](https://github.com/awsdocs/aws-doc-sdk-examples/tree/main/cpp/example_code/glue#code-examples).

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region in which the bucket was created
(overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::Glue::GlueClient client(clientConfig);

        Aws::Glue::Model::StartCrawlerRequest request;
        request.SetName(CRAWLER_NAME);

        Aws::Glue::Model::StartCrawlerOutcome outcome =
client.StartCrawler(request);


        if (outcome.IsSuccess() || (Aws::Glue::GlueErrors::CRAWLER_RUNNING ==
                                    outcome.GetError().GetErrorType())) {
            if (!outcome.IsSuccess()) {
                std::cout << "Crawler was already started." << std::endl;
            }
            else {
                std::cout << "Successfully started crawler." << std::endl;
            }

            std::cout << "This may take a while to run." << std::endl;

            Aws::Glue::Model::CrawlerState crawlerState =
Aws::Glue::Model::CrawlerState::NOT_SET;
            int iterations = 0;
            while (Aws::Glue::Model::CrawlerState::READY != crawlerState) {
                std::this_thread::sleep_for(std::chrono::seconds(1));
                ++iterations;
                if ((iterations % 10) == 0) { // Log status every 10 seconds.
```

```
                    std::cout << "Crawler status " <<
 
Aws::Glue::Model::CrawlerStateMapper::GetNameForCrawlerState(
                                     crawlerState)
                             << ". After " << iterations
                             << " seconds elapsed."
                             << std::endl;
                }
                Aws::Glue::Model::GetCrawlerRequest getCrawlerRequest;
                getCrawlerRequest.SetName(CRAWLER_NAME);

                Aws::Glue::Model::GetCrawlerOutcome getCrawlerOutcome =
client.GetCrawler(
                        getCrawlerRequest);

                if (getCrawlerOutcome.IsSuccess()) {
                    crawlerState =
getCrawlerOutcome.GetResult().GetCrawler().GetState();
                }
                else {
                    std::cerr << "Error getting crawler.  "
                              << getCrawlerOutcome.GetError().GetMessage() <<
std::endl;
                    break;
                }
            }

            if (Aws::Glue::Model::CrawlerState::READY == crawlerState) {
                std::cout << "Crawler finished running after " << iterations
                          << " seconds."
                          << std::endl;
            }
        }
        else {
            std::cerr << "Error starting a crawler.  "
                      << outcome.GetError().GetMessage()
                      << std::endl;

            deleteAssets(CRAWLER_NAME, CRAWLER_DATABASE_NAME, "", bucketName,
                         clientConfig);
            return false;
        }
```

- For API details, see StartCrawler in *AWS SDK for C++ API Reference*.

CLI

AWS CLI

To start a crawler

The following `start-crawler` example starts a crawler.

```
aws glue start-crawler --name my-crawler
```

Output:

```
None
```

For more information, see Defining Crawlers in the *AWS Glue Developer Guide*.

- For API details, see StartCrawler in *AWS CLI Command Reference*.

Java

SDK for Java 2.x

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
    /**
     * Starts a specific AWS Glue crawler.
     *
     * @param glueClient  the AWS Glue client to use for the crawler operation
     * @param crawlerName the name of the crawler to start
     * @throws GlueException if there is an error starting the crawler
     */
    public static void startSpecificCrawler(GlueClient glueClient, String
 crawlerName) {
        try {
            StartCrawlerRequest crawlerRequest = StartCrawlerRequest.builder()
                .name(crawlerName)
                .build();

            glueClient.startCrawler(crawlerRequest);
            System.out.println(crawlerName + " was successfully started!");

        } catch (GlueException e) {
            throw e;
        }
    }
```

- For API details, see StartCrawler in *AWS SDK for Java 2.x API Reference*.

JavaScript

SDK for JavaScript (v3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
const startCrawler = (name) => {
  const client = new GlueClient({});

  const command = new StartCrawlerCommand({
    Name: name,
  });

  return client.send(command);
};
```

- For API details, see StartCrawler in *AWS SDK for JavaScript API Reference*.

Kotlin

SDK for Kotlin

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
suspend fun startSpecificCrawler(crawlerName: String?) {
    val request =
        StartCrawlerRequest {
            name = crawlerName
        }

    GlueClient.fromEnvironment { region = "us-west-2" }.use { glueClient ->
        glueClient.startCrawler(request)
        println("$crawlerName was successfully started.")
    }
}
```

- For API details, see StartCrawler in *AWS SDK for Kotlin API reference*.

PHP

SDK for PHP

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        $crawlerName = "example-crawler-test-" . $uniqid;

        $databaseName = "doc-example-database-$uniqid";

        $glueService->startCrawler($crawlerName);
```

```
    public function startCrawler($crawlerName): Result
    {
        return $this->glueClient->startCrawler([
            'Name' => $crawlerName,
        ]);
    }
```

- For API details, see StartCrawler in *AWS SDK for PHP API Reference*.

Python

SDK for Python (Boto3)

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""

    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 Glue client.
        """
        self.glue_client = glue_client


    def start_crawler(self, name):
        """
        Starts a crawler. The crawler crawls its configured target and creates
        metadata that describes the data it finds in the target data source.

        :param name: The name of the crawler to start.
        """
        try:
            self.glue_client.start_crawler(Name=name)
        except ClientError as err:
            logger.error(
                "Couldn't start crawler %s. Here's why: %s: %s",
                name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise
```

- For API details, see StartCrawler in *AWS SDK for Python (Boto3) API Reference*.

Ruby

SDK for Ruby

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
# The `GlueWrapper` class serves as a wrapper around the AWS Glue API, providing
 a simplified interface for common operations.
# It encapsulates the functionality of the AWS SDK for Glue and provides methods
 for interacting with Glue crawlers, databases, tables, jobs, and S3 resources.
# The class initializes with a Glue client and a logger, allowing it to make API
 calls and log any errors or informational messages.
class GlueWrapper
  def initialize(glue_client, logger)
    @glue_client = glue_client
    @logger = logger
  end

  # Starts a crawler with the specified name.
  #
  # @param name [String] The name of the crawler to start.
  # @return [void]
  def start_crawler(name)
    @glue_client.start_crawler(name: name)
  rescue Aws::Glue::Errors::ServiceError => e
    @logger.error("Glue could not start crawler #{name}: \n#{e.message}")
    raise
  end
```

- For API details, see StartCrawler in *AWS SDK for Ruby API Reference*.

Rust

SDK for Rust

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        let start_crawler =
glue.start_crawler().name(self.crawler()).send().await;

        match start_crawler {
            Ok(_) => Ok(()),
            Err(err) => {
                let glue_err: aws_sdk_glue::Error = err.into();
                match glue_err {
                    aws_sdk_glue::Error::CrawlerRunningException(_) => Ok(()),
                    _ => Err(GlueMvpError::GlueSdk(glue_err)),
                }
            }
        }?;
```

- For API details, see StartCrawler in *AWS SDK for Rust API reference*.

Swift

SDK for Swift

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
import AWSClientRuntime
import AWSGlue

    /// Start running an AWS Glue crawler.
    ///
    /// - Parameters:
    ///   - glueClient: The AWS Glue client to use when starting the crawler.
    ///   - name: The name of the crawler to start running.
    ///
    /// - Returns: `true` if the crawler is started successfully, otherwise
 `false`.
    func startCrawler(glueClient: GlueClient, name: String) async -> Bool {
        do {
            _ = try await glueClient.startCrawler(
                input: StartCrawlerInput(name: name)
            )
        } catch {
            print("*** An unexpected error occurred starting the crawler.")
            return false
        }

        return true
    }
```

- For API details, see StartCrawler in *AWS SDK for Swift API reference*.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `StartJobRun` with an AWS SDK or CLI

The following code examples show how to use `StartJobRun`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Learn the basics

.NET

SDK for .NET

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
    /// <summary>
    /// Start an AWS Glue job run.
    /// </summary>
    /// <param name="jobName">The name of the job.</param>
    /// <returns>A string representing the job run Id.</returns>
    public async Task<string> StartJobRunAsync(
        string jobName,
        string inputDatabase,
        string inputTable,
        string bucketName)
    {
        var request = new StartJobRunRequest
        {
            JobName = jobName,
            Arguments = new Dictionary<string, string>
            {
                {"--input_database", inputDatabase},
                {"--input_table", inputTable},
                {"--output_bucket_url", $"s3://{bucketName}/"}
            }
        };

        var response = await _amazonGlue.StartJobRunAsync(request);
        return response.JobRunId;
    }
```

- For API details, see StartJobRun in *AWS SDK for .NET API Reference*.

C++

SDK for C++

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region in which the bucket was created
(overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::Glue::GlueClient client(clientConfig);

        Aws::Glue::Model::StartJobRunRequest request;
        request.SetJobName(JOB_NAME);

        Aws::Map<Aws::String, Aws::String> arguments;
        arguments["--input_database"] = CRAWLER_DATABASE_NAME;
        arguments["--input_table"] = tableName;
        arguments["--output_bucket_url"] = Aws::String("s3://") + bucketName +
"/";
        request.SetArguments(arguments);

        Aws::Glue::Model::StartJobRunOutcome outcome =
client.StartJobRun(request);

        if (outcome.IsSuccess()) {
            std::cout << "Successfully started the job." << std::endl;

            Aws::String jobRunId = outcome.GetResult().GetJobRunId();
            int iterator = 0;
            bool done = false;
            while (!done) {
                ++iterator;
                std::this_thread::sleep_for(std::chrono::seconds(1));
                Aws::Glue::Model::GetJobRunRequest jobRunRequest;
                jobRunRequest.SetJobName(JOB_NAME);
                jobRunRequest.SetRunId(jobRunId);

                Aws::Glue::Model::GetJobRunOutcome jobRunOutcome =
 client.GetJobRun(
                        jobRunRequest);

                if (jobRunOutcome.IsSuccess()) {
                    const Aws::Glue::Model::JobRun &jobRun =
 jobRunOutcome.GetResult().GetJobRun();
                    Aws::Glue::Model::JobRunState jobRunState =
 jobRun.GetJobRunState();

                    if ((jobRunState == Aws::Glue::Model::JobRunState::STOPPED)
 ||
                        (jobRunState == Aws::Glue::Model::JobRunState::FAILED) ||
                        (jobRunState == Aws::Glue::Model::JobRunState::TIMEOUT))
 {
                        std::cerr << "Error running job. "
                                  << jobRun.GetErrorMessage()
                                  << std::endl;
                        deleteAssets(CRAWLER_NAME, CRAWLER_DATABASE_NAME,
 JOB_NAME,
                                     bucketName,
                                     clientConfig);
                        return false;
                    }
                    else if (jobRunState ==
                             Aws::Glue::Model::JobRunState::SUCCEEDED) {
                        std::cout << "Job run succeeded after  " << iterator <<
                                  " seconds elapsed." << std::endl;
                        done = true;
                    }
                    else if ((iterator % 10) == 0) { // Log status every 10
 seconds.
                        std::cout << "Job run status " <<

 Aws::Glue::Model::JobRunStateMapper::GetNameForJobRunState(
                                          jobRunState) <<
                                  ". " << iterator <<
                                  " seconds elapsed." << std::endl;
                    }
                }
                else {
                    std::cerr << "Error retrieving job run state. "
                              << jobRunOutcome.GetError().GetMessage()
                              << std::endl;
                    deleteAssets(CRAWLER_NAME, CRAWLER_DATABASE_NAME, JOB_NAME,
                                 bucketName, clientConfig);
                    return false;
                }
            }
        }
        else {
            std::cerr << "Error starting a job. " <<
outcome.GetError().GetMessage()
                      << std::endl;
            deleteAssets(CRAWLER_NAME, CRAWLER_DATABASE_NAME, JOB_NAME,
bucketName,
                         clientConfig);
            return false;
        }
```

- For API details, see StartJobRun in *AWS SDK for C++ API Reference*.

CLI

AWS CLI

To start running a job

The following `start-job-run` example starts a job.

```
aws glue start-job-run \
    --job-name my-job
```

Output:

```
{
    "JobRunId":
 "jr_22208b1f44eb5376a60569d4b21dd20fcb8621e1a366b4e7b2494af764b82ded"
}
```

For more information, see Authoring Jobs in the *AWS Glue Developer Guide*.

- For API details, see StartJobRun in *AWS CLI Command Reference*.

Java

SDK for Java 2.x

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
    /**
     * Starts a job run in AWS Glue.
     *
     * @param glueClient    the AWS Glue client to use for the job run
     * @param jobName       the name of the Glue job to run
     * @param inputDatabase the name of the input database
     * @param inputTable    the name of the input table
     * @param outBucket     the URL of the output S3 bucket
     * @throws GlueException if there is an error starting the job run
     */
    public static void startJob(GlueClient glueClient, String jobName, String
 inputDatabase, String inputTable,
                               String outBucket) {
        try {
            Map<String, String> myMap = new HashMap<>();
            myMap.put("--input_database", inputDatabase);
            myMap.put("--input_table", inputTable);
            myMap.put("--output_bucket_url", outBucket);

            StartJobRunRequest runRequest = StartJobRunRequest.builder()
                .workerType(WorkerType.G_1_X)
                .numberOfWorkers(10)
                .arguments(myMap)
                .jobName(jobName)
                .build();

            StartJobRunResponse response = glueClient.startJobRun(runRequest);
            System.out.println("The request Id of the job is " +
 response.responseMetadata().requestId());

        } catch (GlueException e) {
            throw e;
        }
    }
```

- For API details, see StartJobRun in *AWS SDK for Java 2.x API Reference*.

JavaScript

SDK for JavaScript (v3)

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
const startJobRun = (jobName, dbName, tableName, bucketName) => {
  const client = new GlueClient({});

  const command = new StartJobRunCommand({
    JobName: jobName,
    Arguments: {
      "--input_database": dbName,
      "--input_table": tableName,
      "--output_bucket_url": `s3://${bucketName}/`,
    },
  });

  return client.send(command);
};
```

- For API details, see StartJobRun in *AWS SDK for JavaScript API Reference*.

PHP

SDK for PHP

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        $jobName = 'test-job-' . $uniqid;

        $databaseName = "doc-example-database-$uniqid";

        $tables = $glueService->getTables($databaseName);

        $outputBucketUrl = "s3://$bucketName";
        $runId = $glueService->startJobRun($jobName, $databaseName, $tables,
$outputBucketUrl)['JobRunId'];

    public function startJobRun($jobName, $databaseName, $tables,
$outputBucketUrl): Result
    {
        return $this->glueClient->startJobRun([
            'JobName' => $jobName,
            'Arguments' => [
                'input_database' => $databaseName,
                'input_table' => $tables['TableList'][0]['Name'],
                'output_bucket_url' => $outputBucketUrl,
                '--input_database' => $databaseName,
                '--input_table' => $tables['TableList'][0]['Name'],
                '--output_bucket_url' => $outputBucketUrl,
            ],
        ]);
    }
```

- For API details, see StartJobRun in *AWS SDK for PHP API Reference*.

Python

SDK for Python (Boto3)

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the [AWS Code Examples Repository](https://github.com/awsdocs/aws-doc-sdk-examples/tree/main/python/example_code/glue#code-examples).

```
class GlueWrapper:
    """Encapsulates AWS Glue actions."""

    def __init__(self, glue_client):
        """
        :param glue_client: A Boto3 Glue client.
        """
        self.glue_client = glue_client


    def start_job_run(self, name, input_database, input_table,
 output_bucket_name):
        """
        Starts a job run. A job run extracts data from the source, transforms it,
        and loads it to the output bucket.

        :param name: The name of the job definition.
        :param input_database: The name of the metadata database that contains
 tables
                               that describe the source data. This is typically
 created
                               by a crawler.
        :param input_table: The name of the table in the metadata database that
                            describes the source data.
        :param output_bucket_name: The S3 bucket where the output is written.
        :return: The ID of the job run.
        """
        try:
            # The custom Arguments that are passed to this function are used by
 the
            # Python ETL script to determine the location of input and output
 data.
```

```
            response = self.glue_client.start_job_run(
                JobName=name,
                Arguments={
                    "--input_database": input_database,
                    "--input_table": input_table,
                    "--output_bucket_url": f"s3://{output_bucket_name}/",
                },
            )
        except ClientError as err:
            logger.error(
                "Couldn't start job run %s. Here's why: %s: %s",
                name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise
        else:
            return response["JobRunId"]
```

- For API details, see StartJobRun in *AWS SDK for Python (Boto3) API Reference*.

Ruby

SDK for Ruby

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
# The `GlueWrapper` class serves as a wrapper around the AWS Glue API, providing
 a simplified interface for common operations.
# It encapsulates the functionality of the AWS SDK for Glue and provides methods
 for interacting with Glue crawlers, databases, tables, jobs, and S3 resources.
# The class initializes with a Glue client and a logger, allowing it to make API
 calls and log any errors or informational messages.
class GlueWrapper
  def initialize(glue_client, logger)
```

```
    @glue_client = glue_client
    @logger = logger
  end

  # Starts a job run for the specified job.
  #
  # @param name [String] The name of the job to start the run for.
  # @param input_database [String] The name of the input database for the job.
  # @param input_table [String] The name of the input table for the job.
  # @param output_bucket_name [String] The name of the output S3 bucket for the
 job.
  # @return [String] The ID of the started job run.
  def start_job_run(name, input_database, input_table, output_bucket_name)
    response = @glue_client.start_job_run(
      job_name: name,
      arguments: {
        '--input_database': input_database,
        '--input_table': input_table,
        '--output_bucket_url': "s3://#{output_bucket_name}/"
      }
    )
    response.job_run_id
  rescue Aws::Glue::Errors::GlueException => e
    @logger.error("Glue could not start job run #{name}: \n#{e.message}")
    raise
  end
```

- For API details, see StartJobRun in *AWS SDK for Ruby API Reference*.

Rust

SDK for Rust

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
        let job_run_output = glue
            .start_job_run()
            .job_name(self.job())
            .arguments("--input_database", self.database())
            .arguments(
                "--input_table",
                self.tables
                    .first()
                    .ok_or_else(|| GlueMvpError::Unknown("Missing crawler 
 table".into()))?
                    .name(),
            )
            .arguments("--output_bucket_url", self.bucket())
            .send()
            .await
            .map_err(GlueMvpError::from_glue_sdk)?;

        let job = job_run_output
            .job_run_id()
            .ok_or_else(|| GlueMvpError::Unknown("Missing run id from just 
 started job".into()))?
            .to_string();
```

- For API details, see StartJobRun in *AWS SDK for Rust API reference*.

Swift

SDK for Swift

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
import AWSClientRuntime
import AWSGlue

    /// Start an AWS Glue job run.
    ///
    /// - Parameters:
    ///   - glueClient: The AWS Glue client to use.
    ///   - jobName: The name of the job to run.
    ///   - databaseName: The name of the AWS Glue database to run the job
 against.
    ///   - tableName: The name of the table in the database to run the job
 against.
    ///   - outputURL: The AWS S3 URI of the bucket location into which to
    ///     write the resulting output.
    ///
    /// - Returns: `true` if the job run is started successfully, otherwise
 `false`.
    func startJobRun(glueClient: GlueClient, name jobName: String, databaseName:
 String,
                     tableName: String, outputURL: String) async -> String? {
        do {
            let output = try await glueClient.startJobRun(
                input: StartJobRunInput(
                    arguments: [
                        "--input_database": databaseName,
                        "--input_table": tableName,
                        "--output_bucket_url": outputURL
                    ],
                    jobName: jobName,
                    numberOfWorkers: 10,
                    workerType: .g1x
                )
            )

            guard let id = output.jobRunId else {
                return nil
            }

            return id
        } catch {
            return nil
        }
    }
```

- For API details, see StartJobRun in *AWS SDK for Swift API reference*.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Security in AWS Glue

Cloud security at AWS is the highest priority. As an AWS customer, you benefit from a data center and network architecture that is built to meet the requirements of the most security-sensitive organizations.

Security is a shared responsibility between AWS and you. The shared responsibility model describes this as security *of* the cloud and security *in* the cloud:

- **Security of the cloud** – AWS is responsible for protecting the infrastructure that runs AWS services in the AWS Cloud. AWS also provides you with services that you can use securely. Third-party auditors regularly test and verify the effectiveness of our security as part of the AWS compliance programs. To learn about the compliance programs that apply to AWS Glue, see AWS Services in Scope by Compliance Program.
- **Security in the cloud** – Your responsibility is determined by the AWS service that you use. You are also responsible for other factors including the sensitivity of your data, your company's requirements, and applicable laws and regulations.

This documentation helps you understand how to apply the shared responsibility model when using AWS Glue. The following topics show you how to configure AWS Glue to meet your security and compliance objectives. You also learn how to use other AWS services that help you to monitor and secure your AWS Glue resources.

Topics

- Data protection in AWS Glue
- Identity and access management for AWS Glue
- AWS Lake Formation access control models
- Using Amazon S3 Access Grants with AWS Glue
- Trusted Identity Propagation with AWS Glue ETL
- Logging and monitoring in AWS Glue
- Compliance validation for AWS Glue
- Resilience in AWS Glue
- Infrastructure security in AWS Glue

Data protection in AWS Glue

AWS Glue offers several features that are designed to help protect your data.

Topics

- Encrypting data at rest
- Encrypting data in transit
- FIPS compliance
- Key management
- AWS Glue dependency on other AWS services
- Development endpoints

Encrypting data at rest

AWS Glue supports data encryption at rest for Building visual ETL jobs and Developing scripts using development endpoints. You can configure extract, transform, and load (ETL) jobs and development endpoints to use AWS Key Management Service (AWS KMS) keys to write encrypted data at rest. You can also encrypt the metadata stored in the AWS Glue Data Catalog using keys that you manage with AWS KMS. Additionally, you can use AWS KMS keys to encrypt job bookmarks and the logs generated by crawlers and ETL jobs.

You can encrypt metadata objects in your AWS Glue Data Catalog in addition to the data written to Amazon Simple Storage Service (Amazon S3) and Amazon CloudWatch Logs by jobs, crawlers, and development endpoints. When you create jobs, crawlers, and development endpoints in AWS Glue, you can provide encryption settings by attaching a security configuration. Security configurations contain Amazon S3-managed server-side encryption keys (SSE-S3) or customer master keys (CMKs) stored in AWS KMS (SSE-KMS). You can create security configurations using the AWS Glue console.

You can also turn on encryption of the entire Data Catalog in your account. You do so by specifying CMKs stored in AWS KMS.

> ⚠ **Important**
>
> AWS Glue supports only symmetric customer managed keys. For more information, see Customer Managed Keys (CMKs) in the *AWS Key Management Service Developer Guide*.

With encryption turned on, when you add Data Catalog objects, run crawlers, run jobs, or start development endpoints, SSE-S3 or SSE-KMS keys are used to write data at rest. In addition, you can configure AWS Glue to only access Java Database Connectivity (JDBC) data stores through a trusted Transport Layer Security (TLS) protocol.

In AWS Glue, you control encryption settings in the following places:

- The settings of your Data Catalog.
- The security configurations that you create.
- The server-side encryption setting (SSE-S3 or SSE-KMS) that is passed as a parameter to your AWS Glue ETL (extract, transform, and load) job.

For more information about how to set up encryption, see Setting up encryption in AWS Glue.

Topics

- Encrypting your Data Catalog
- Encrypting connection passwords
- Encrypting data written by AWS Glue

Encrypting your Data Catalog

AWS Glue Data Catalog encryption provides enhanced security for your sensitive data. AWS Glue integrates with AWS Key Management Service (AWS KMS) to encrypt metadata that's stored in the Data Catalog. You can enable or disable encryption settings for resources in the Data Catalog using the AWS Glue console or the AWS CLI.

When you enable encryption for your Data Catalog, all new objects that you create will be encrypted. When you disable encryption, the new objects you create will not be encrypted, but existing encrypted objects will remain encrypted.

You can encrypt your entire Data Catalog using AWS managed encryption keys or customer managed encryption keys. For more information on key types and states, see AWS Key Management Service concepts in the AWS Key Management Service Developer Guide.

> **Note**
>
> When you use the encrypted Data Catalog with a crawler, you must maintain the encryption settings. Removing encryption settings after a crawler has processed an

encrypted catalog results in errors. If you need to remove encryption settings, create a new crawler instead of modifying the existing one.

AWS managed keys

AWS managed keys are KMS keys in your account that are created, managed, and used on your behalf by an AWS service that's integrated with AWS KMS. You can view the AWS managed keys in your account, view their key policies, and audit their use in AWS CloudTrail logs. However, you can't manage these keys or change their permissions.

Encryption at rest automatically integrates with AWS KMS for managing the AWS managed keys for AWS Glue that are used to encrypt your metadata. If an AWS managed key doesn't exist when you enable metadata encryption, AWS KMS automatically creates a new key for you.

For more information, see AWS managed keys.

Customer managed keys

Customer managed keys are KMS keys in your AWS account that you create, own, and manage. You have full control over these KMS keys. You can:

- Establish and maintain their key policies, IAM policies, and grants
- Enable and disable them
- Rotate their cryptographic material
- Add tags
- Create aliases that refer to them
- Schedule them for deletion

For more information about managing the permissions of a customer managed key, see Customer managed keys.

Important

AWS Glue supports only symmetric customer managed keys. The KMS key list displays only symmetric keys. However, if you select **Choose a KMS key ARN**, the console lets you enter an ARN for any key type. Ensure that you enter only ARNs for symmetric keys.

To create a symmetric customer managed key, follow the steps for creating symmetric customer managed keys in the AWS Key Management Service Developer Guide.

When you enable Data Catalog encryption at rest, the following resource types are encrypted using KMS keys:

- Databases
- Tables
- Partitions
- Table versions
- Column statistics
- User-defined functions
- Data Catalog views

AWS Glue encryption context

An encryption context is an optional set of key-value pairs that contain additional contextual information about the data. AWS KMS uses the encryption context as additional authenticated data to support authenticated encryption. When you include an encryption context in a request to encrypt data, AWS KMS binds the encryption context to the encrypted data. To decrypt data, you include the same encryption context in the request. AWS Glue uses the same encryption context in all AWS KMS cryptographic operations, where the key is `glue_catalog_id` and the value is the `catalogId`.

```
"encryptionContext": {
    "glue_catalog_id": "111122223333"
}
```

When you use an AWS managed key or a symmetric customer managed key to encrypt your Data Catalog, you can also use the encryption context in audit records and logs to identify how the key is being used. The encryption context also appears in logs that are generated by AWS CloudTrail or Amazon CloudWatch logs.

Enabling encryption

You can enable encryption for your AWS Glue Data Catalog objects in the **Data Catalog settings** in the AWS Glue console or by using the AWS CLI.

Console

To enable encryption using the console

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. Choose **Data Catalog** in the navigation pane.
3. On the **Data Catalog settings** page, select the **Metadata encryption** check box, and choose an AWS KMS key.

   When you enable encryption, if you don’t specify a customer managed key, the encryption settings use an AWS managed KMS key.
4. (Optional) When you use a customer managed key to encrypt your Data Catalog, the Data Catalog provides an option to register an IAM role to encrypt and decrypt resources. You need to grant your IAM role permissions that AWS Glue can assume on your behalf. This includes AWS KMS permissions to encrypt and decrypt data.

   When you create a new resource in the Data Catalog, AWS Glue assumes the IAM role that's provided to encrypt the data. Similarly, when a consumer accesses the resource, AWS Glue assumes the IAM role to decrypt data. If you register an IAM role with the required permissions, the calling principal no longer requires permissions to access the key and decrypt the data.

   > ⚠ **Important**
   >
   > You can delegate KMS operations to an IAM role only when you use a customer managed key to encrypt the Data Catalog resources. KMS role delegation feature doesn't support using AWS managed keys for encrypting Data Catalog resources at this time.

⚠ **Warning**

When you enable an IAM role to delegate KMS operations, you can no longer access the Data Catalog resources that were encrypted previously with an AWS managed key.

a. To enable an IAM role that AWS Glue can assume to encrypt and decrypt data on your behalf, select the **Delegate KMS operations to an IAM role** option.

b. Next, choose an IAM role.

   To create an IAM role, see [Create an IAM role for AWS Glue](#).

   The IAM role that AWS Glue assumes to access the Data Catalog must have the permissions to encrypt and decrypt metadata in the Data Catalog. You can create an IAM role, and attach the following inline policies:

   - Add the following policy to include AWS KMS permissions to encrypt and decrypt the Data Catalog.

     JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "kms:Decrypt",
        "kms:Encrypt",
        "kms:GenerateDataKey"
      ],
      "Resource": "arn:aws:kms:us-east-1:111122223333:key/<key-id>"
    }
  ]
}
```

   - Next, add the following trust policy to the role for AWS Glue service to assume the IAM role.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "",
      "Effect": "Allow",
      "Principal": {
        "Service": "glue.amazonaws.com"
      },
      "Action": "sts:AssumeRole"
    }
  ]
}
```

- Next, add the `iam:PassRole` permission to the IAM role.

  JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "iam:PassRole"
      ],
      "Resource": [
        "arn:aws:iam::111122223333:role/<encryption-role-name>"
      ]
    }
  ]
}
```

When you enable encryption, if you haven't specified an IAM role for AWS Glue to assume, the principal accessing the Data Catalog must have permissions to perform the following API operations:

- `kms:Decrypt`
- `kms:Encrypt`

- `kms:GenerateDataKey`

AWS CLI

To enable encryption using the SDK or AWS CLI

- Use the `PutDataCatalogEncryptionSettings` API operation. If no key is specified, AWS Glue uses AWS managed encryption key for the customer account to encrypt the Data Catalog.

```
aws glue put-data-catalog-encryption-settings \
  --data-catalog-encryption-settings '{
      "EncryptionAtRest": {
      "CatalogEncryptionMode": "SSE-KMS-WITH-SERVICE-ROLE",
      "SseAwsKmsKeyId": "arn:aws:kms:<region>:<account-id>:key/<key-id>",
      "CatalogEncryptionServiceRole":"arn:aws:iam::<account-
id>:role/<encryption-role-name>"
    }

  }'
```

When you enable encryption, all objects that you create in the Data Catalog objects are encrypted. If you clear this setting, the objects you create in the Data Catalog are no longer encrypted. You can continue to access the existing encrypted objects in the Data Catalog with the required KMS permissions.

> ⚠ **Important**
>
> The AWS KMS key must remain available in the AWS KMS key store for any objects that are encrypted with it in the Data Catalog. If you remove the key, the objects can no longer be decrypted. You might want this in some scenarios to prevent access to Data Catalog metadata.

Monitoring your KMS keys for AWS Glue

When you use KMS keys with your Data Catalog resources, you can use AWS CloudTrail or Amazon CloudWatch Logs to track requests that AWS Glue sends to AWS KMS. AWS CloudTrail monitors and records KMS operations that AWS Glue calls to access data that's encrypted by your KMS keys.

The following examples are AWS CloudTrail events for the `Decrypt` and `GenerateDataKey` operations.

Decrypt

```
{
    "eventVersion": "1.08",
    "userIdentity": {
        "type": "AssumedRole",
        "principalId": "AROAXPHTESTANDEXAMPLE:Sampleuser01",
        "arn": "arn:aws:sts::111122223333:assumed-role/Admin/Sampleuser01",
        "accountId": "111122223333",
        "accessKeyId": "AKIAIOSFODNN7EXAMPLE",
        "sessionContext": {
            "sessionIssuer": {
                "type": "Role",
                "principalId": "AROAXPHTESTANDEXAMPLE",
                "arn": "arn:aws:iam::111122223333:role/Admin",
                "accountId": "111122223333",
                "userName": "Admin"
            },
            "webIdFederationData": {},
            "attributes": {
                "creationDate": "2024-01-10T14:33:56Z",
                "mfaAuthenticated": "false"
            }
        },
        "invokedBy": "glue.amazonaws.com"
    },
    "eventTime": "2024-01-10T15:18:11Z",
    "eventSource": "kms.amazonaws.com",
    "eventName": "Decrypt",
    "awsRegion": "eu-west-2",
    "sourceIPAddress": "glue.amazonaws.com",
    "userAgent": "glue.amazonaws.com",
    "requestParameters": {
        "encryptionContext": {
            "glue_catalog_id": "111122223333"
        },
        "encryptionAlgorithm": "SYMMETRIC_DEFAULT"
    },
    "responseElements": null,
    "requestID": "43b019aa-34b8-4798-9b98-ee968b2d63df",
    "eventID": "d7614763-d3fe-4f84-a1e1-3ca4d2a5bbd5",
    "readOnly": true,
    "resources": [
        {
            "accountId": "111122223333",
            "type": "AWS::KMS::Key",
            "ARN": "arn:aws:kms:<region>:111122223333:key/<key-id>"
        }
    ],
    "eventType": "AwsApiCall",
    "managementEvent": true,
    "recipientAccountId": "111122223333",
    "eventCategory": "Management",
    "sessionCredentialFromConsole": "true"
}
```

GenerateDataKey

```
{
    "eventVersion": "1.08",
    "userIdentity": {
        "type": "AssumedRole",
        "principalId":
 "AROAXPHTESTANDEXAMPLE:V_00_GLUE_KMS_GENERATE_DATA_KEY_111122223333",
        "arn": "arn:aws:sts::111122223333:assumed-role/Admin/
V_00_GLUE_KMS_GENERATE_DATA_KEY_111122223333",
        "accountId": "111122223333",
        "accessKeyId": "AKIAIOSFODNN7EXAMPLE",
        "sessionContext": {
            "sessionIssuer": {
                "type": "Role",
                "principalId": "AROAXPHTESTANDEXAMPLE",
                "arn": "arn:aws:iam::111122223333:role/Admin",
                "accountId": "AKIAIOSFODNN7EXAMPLE",
                "userName": "Admin"
            },
            "webIdFederationData": {},
            "attributes": {
                "creationDate": "2024-01-05T21:15:47Z",
                "mfaAuthenticated": "false"
            }
        },
        "invokedBy": "glue.amazonaws.com"
    },
    "eventTime": "2024-01-05T21:15:47Z",
    "eventSource": "kms.amazonaws.com",
    "eventName": "GenerateDataKey",
    "awsRegion": "eu-west-2",
    "sourceIPAddress": "glue.amazonaws.com",
    "userAgent": "glue.amazonaws.com",
    "requestParameters": {
        "keyId": "arn:aws:kms:eu-west-2:AKIAIOSFODNN7EXAMPLE:key/
AKIAIOSFODNN7EXAMPLE",
        "encryptionContext": {
            "glue_catalog_id": "111122223333"
        },
        "keySpec": "AES_256"
    },
    "responseElements": null,
    "requestID": "64d1783a-4b62-44ba-b0ab-388b50188070",
    "eventID": "1c73689b-2ef2-443b-aed7-8c126585ca5e",
    "readOnly": true,
    "resources": [
        {
            "accountId": "111122223333",
            "type": "AWS::KMS::Key",
            "ARN": "arn:aws:kms:eu-west-2:111122223333:key/AKIAIOSFODNN7EXAMPLE"
        }
    ],
    "eventType": "AwsApiCall",
    "managementEvent": true,
    "recipientAccountId": "111122223333",
    "eventCategory": "Management"
}
```

Encrypting connection passwords

You can retrieve connection passwords in the AWS Glue Data Catalog by using the `GetConnection` and `GetConnections` API operations. These passwords are stored in the Data Catalog connection and are used when AWS Glue connects to a Java Database Connectivity (JDBC) data store. When the connection was created or updated, an option in the Data Catalog settings determined whether the password was encrypted, and if so, what AWS Key Management Service (AWS KMS) key was specified.

On the AWS Glue console, you can turn on this option on the **Data catalog settings** page.

To encrypt connection passwords

1. Sign in to the AWS Management Console and open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. Choose **Settings** in the navigation pane.
3. On the **Data catalog settings** page, select **Encrypt connection passwords**, and choose an AWS KMS key.

   > ⚠ **Important**
   >
   > AWS Glue supports only symmetric customer master keys (CMKs). The **AWS KMS key** list displays only symmetric keys. However, if you select **Choose a AWS KMS key ARN**, the console lets you enter an ARN for any key type. Ensure that you enter only ARNs for symmetric keys.

   For more information, see Data Catalog settings .

Encrypting data written by AWS Glue

A *security configuration* is a set of security properties that can be used by AWS Glue. You can use a security configuration to encrypt data at rest. The following scenarios show some of the ways that you can use a security configuration.

- Attach a security configuration to an AWS Glue crawler to write encrypted Amazon CloudWatch Logs. For more information about attaching security configurations to crawlers, see the section called “Configure security settings”.
- Attach a security configuration to an extract, transform, and load (ETL) job to write encrypted Amazon Simple Storage Service (Amazon S3) targets and encrypted CloudWatch Logs.
- Attach a security configuration to an ETL job to write its jobs bookmarks as encrypted Amazon S3 data.
- Attach a security configuration to a development endpoint to write encrypted Amazon S3 targets.

> ⚠ **Important**
>
> Currently, a security configuration overrides any server-side encryption (SSE-S3) setting that is passed as an ETL job parameter. Thus, if both a security configuration and an SSE-S3 parameter are associated with a job, the SSE-S3 parameter is ignored.

For more information about security configurations, see Managing security configurations on the AWS Glue console.

Topics

- Setting Up AWS Glue to use security configurations
- Creating a route to AWS KMS for VPC jobs and crawlers
- Managing security configurations on the AWS Glue console

Setting Up AWS Glue to use security configurations

Follow these steps to set up your AWS Glue environment to use security configurations.

1. Create or update your AWS Key Management Service (AWS KMS) keys to grant AWS KMS permissions to the IAM roles that are passed to AWS Glue crawlers and jobs to encrypt CloudWatch Logs. For more information, see Encrypt Log Data in CloudWatch Logs Using AWS KMS in the *Amazon CloudWatch Logs User Guide*.

   In the following example, `"role1"`, `"role2"`, and `"role3"` are IAM roles that are passed to crawlers and jobs.

```
{
        "Effect": "Allow",
        "Principal": { "Service": "logs.region.amazonaws.com",
        "AWS": [
                  "role1",
                  "role2",
                  "role3"
              ] },
                     "Action": [
                            "kms:Encrypt*",
                            "kms:Decrypt*",
                            "kms:ReEncrypt*",
                            "kms:GenerateDataKey*",
                            "kms:Describe*"
                     ],
                     "Resource": "*"
}
```

The `Service` statement, shown as `"Service": "logs.region.amazonaws.com"`, is required if you use the key to encrypt CloudWatch Logs.

2. Ensure that the AWS KMS key is `ENABLED` before it is used.

Note

If you are using Iceberg as your data lake framework, Iceberg tables have their own mechanisms to enable server-side encryption. You should enable these configuration in addition to AWS Glue's security configurations. To enable server-side encryption on Iceberg tables, review the guidance from Iceberg documentation.

Creating a route to AWS KMS for VPC jobs and crawlers

You can connect directly to AWS KMS through a private endpoint in your virtual private cloud (VPC) instead of connecting over the internet. When you use a VPC endpoint, communication between your VPC and AWS KMS is conducted entirely within the AWS network.

You can create an AWS KMS VPC endpoint within a VPC. Without this step, your jobs or crawlers might fail with a `kms timeout` on jobs or an `internal service exception` on crawlers.

For detailed instructions, see Connecting to AWS KMS Through a VPC Endpoint in the *AWS Key Management Service Developer Guide*.

As you follow these instructions, on the VPC console, you must do the following:

- Select **Enable Private DNS name**.
- Choose the **Security group** (with self-referencing rule) that you use for your job or crawler that accesses Java Database Connectivity (JDBC). For more information about AWS Glue connections, see Connecting to data.

When you add a security configuration to a crawler or job that accesses JDBC data stores, AWS Glue must have a route to the AWS KMS endpoint. You can provide the route with a network address translation (NAT) gateway or with an AWS KMS VPC endpoint. To create a NAT gateway, see NAT Gateways in the *Amazon VPC User Guide*.

Managing security configurations on the AWS Glue console

> ⚠ **Warning**
>
> AWS Glue security configurations are not currently supported in Ray jobs.

A *security configuration* in AWS Glue contains the properties that are needed when you write encrypted data. You create security configurations on the AWS Glue console to provide the encryption properties that are used by crawlers, jobs, and development endpoints.

To see a list of all the security configurations that you have created, open the AWS Glue console at https://console.aws.amazon.com/glue/ and choose **Security configurations** in the navigation pane.

The **Security configurations** list displays the following properties about each configuration:

Name

The unique name you provided when you created the configuration. The name may contain letters (A-Z), numbers (0-9), hypens (-), or underscores (_), and be up to 255 characters long.

Enable Amazon S3 encryption

If turned on, the Amazon Simple Storage Service (Amazon S3) encryption mode such as `SSE-KMS` or `SSE-S3` is enabled for metadata store in the data catalog.

Enable Amazon CloudWatch logs encryption

If turned on, the Amazon S3 encryption mode such as `SSE-KMS` is enabled when writing logs to Amazon CloudWatch.

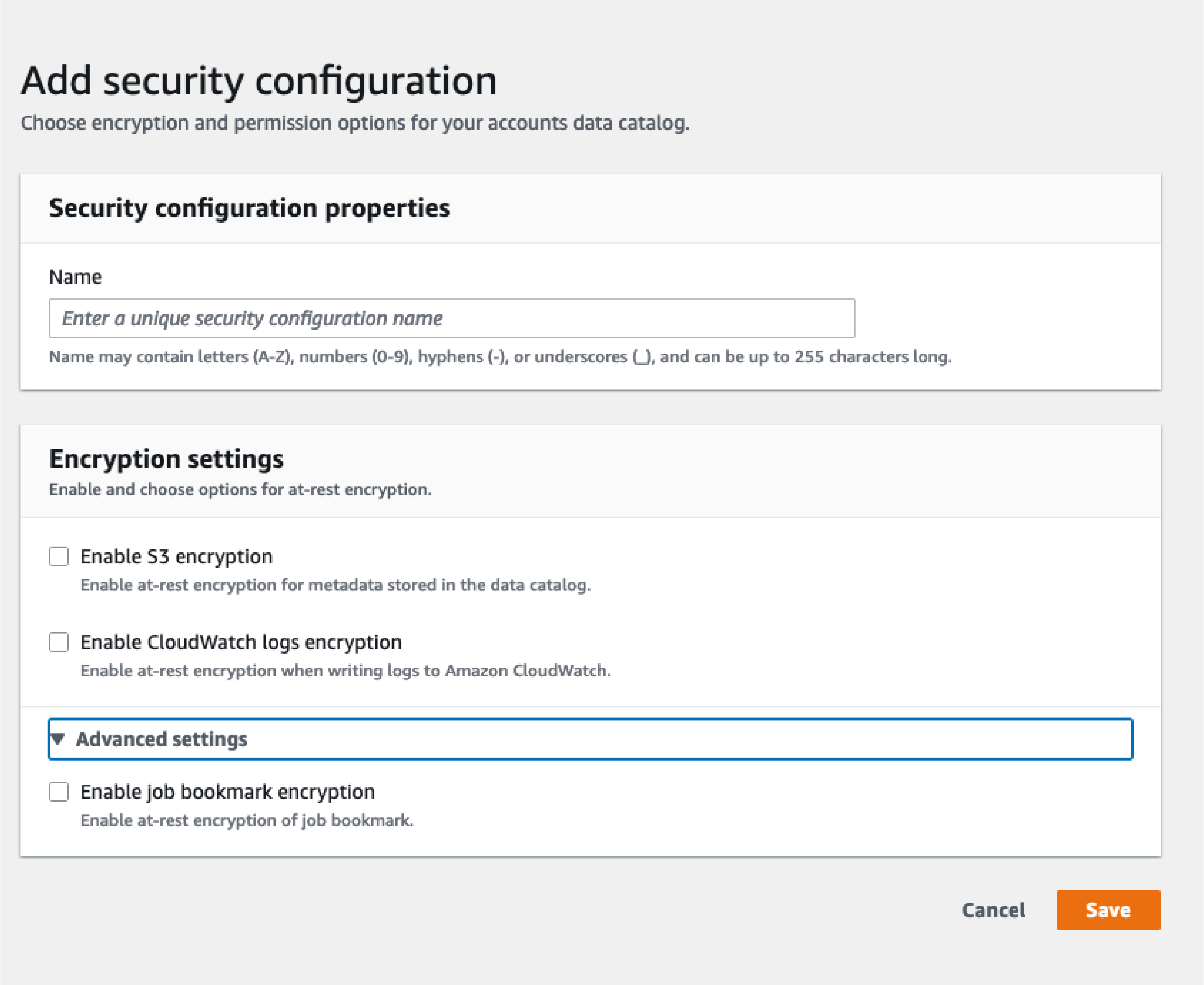
Advanced settings: Enable job bookmark encryption

If turned on, the Amazon S3 encryption mode such as `CSE-KMS` is enabled when jobs are bookmarked.

You can add or delete configurations in the **Security configurations** section on the console. To see more details for a configuration, choose the configuration name in the list. Details include the information that you defined when you created the configuration.

Adding a security configuration

To add a security configuration using the AWS Glue console, on the **Security configurations** page, choose **Add security configuration**.



Add security configuration

Choose encryption and permission options for your accounts data catalog.

Security configuration properties

Name

Enter a unique security configuration name

Name may contain letters (A-Z), numbers (0-9), hyphens (-), or underscores (_), and can be up to 255 characters long.

Encryption settings

Enable and choose options for at-rest encryption.

☐ Enable S3 encryption

Enable at-rest encryption for metadata stored in the data catalog.

☐ Enable CloudWatch logs encryption

Enable at-rest encryption when writing logs to Amazon CloudWatch.

▼ Advanced settings

☐ Enable job bookmark encryption

Enable at-rest encryption of job bookmark.

Security configuration properties

Enter a unique security configuration name. The name may contain letters (A-Z), numbers (0-9), hyphens (-), or underscores (_), and can be up to 255 characters long.

Encryption settings

You can enable at-rest encryption for metadata stored in the Data Catalog in Amazon S3 and logs in Amazon CloudWatch. To set up encryption of data and metadata with AWS Key Management Service (AWS KMS) keys on the AWS Glue console, add a policy to the console user. This policy must specify the allowed resources as key Amazon Resource Names (ARNs) that are used to encrypt Amazon S3 data stores, as in the following example.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": {
    "Effect": "Allow",
    "Action": [
      "kms:GenerateDataKey",
      "kms:Decrypt",
      "kms:Encrypt"
    ],
    "Resource": "arn:aws:kms:us-east-1:111122223333:key/key-id"
  }
}
```

> ⚠ **Important**
>
> When a security configuration is attached to a crawler or job, the IAM role that is passed must have AWS KMS permissions. For more information, see Encrypting data written by AWS Glue.

When you define a configuration, you can provide values for the following properties:

Enable S3 encryption

When you are writing Amazon S3 data, you use either server-side encryption with Amazon S3 managed keys (SSE-S3) or server-side encryption with AWS KMS managed keys (SSE-KMS). This field is optional. To allow access to Amazon S3, choose an AWS KMS key, or choose **Enter a key ARN** and provide the ARN for the key. Enter the ARN in the form `arn:aws:kms:region:account-id:key/key-id`. You can also provide the ARN as a key alias, such as `arn:aws:kms:region:account-id:alias/alias-name`.

If you enable Spark UI for your job, the Spark UI log file uploaded to Amazon S3 will be applied with the same encryption.

> ⚠ **Important**
>
> AWS Glue supports only symmetric customer master keys (CMKs). The **AWS KMS key** list displays only symmetric keys. However, if you select **Choose a AWS KMS key ARN**, the console lets you enter an ARN for any key type. Ensure that you enter only ARNs for symmetric keys.

Enable CloudWatch Logs encryption

Server-side (SSE-KMS) encryption is used to encrypt CloudWatch Logs. This field is optional. To turn it on, choose an AWS KMS key, or choose **Enter a key ARN** and provide the ARN for the key. Enter the ARN in the form `arn:aws:kms:region:account-id:key/key-id`. You can also provide the ARN as a key alias, such as `arn:aws:kms:region:account-id:alias/alias-name`.

Advanced settings: Job bookmark encryption

Client-side (CSE-KMS) encryption is used to encrypt job bookmarks. This field is optional. The bookmark data is encrypted before it is sent to Amazon S3 for storage. To turn it on, choose an AWS KMS key, or choose **Enter a key ARN** and provide the ARN for the key. Enter the ARN in the form `arn:aws:kms:region:account-id:key/key-id`. You can also provide the ARN as a key alias, such as `arn:aws:kms:region:account-id:alias/alias-name`.

For more information, see the following topics in the *Amazon Simple Storage Service User Guide*:

- For information about `SSE-S3`, see Protecting Data Using Server-Side Encryption with Amazon S3-Managed Encryption Keys (SSE-S3).
- For information about `SSE-KMS`, see Protecting Data Using Server-Side Encryption with AWS KMS keys.
- For information about `CSE-KMS`, see Using a KMS key stored in AWS KMS.

Encrypting data in transit

AWS provides Transport Layer Security (TLS) encryption for data in motion. You can configure encryption settings for crawlers, ETL jobs, and development endpoints using security configurations in AWS Glue. You can turn on AWS Glue Data Catalog encryption via the settings for the Data Catalog.

As of September 4, 2018, AWS KMS (*bring your own key* and *server-side encryption*) for AWS Glue ETL and the AWS Glue Data Catalog is supported.

FIPS compliance

If you require FIPS 140-2 validated cryptographic modules when accessing AWS through a command line interface or an API, use a FIPS endpoint. For more information about the available FIPS endpoints, see Federal Information Processing Standard (FIPS) 140-2.

Key management

You can use AWS Identity and Access Management (IAM) with AWS Glue to define users, AWS resources, groups, roles and fine-grained policies regarding access, denial, and more.

You can define the access to the metadata using both resource-based and identity-based policies, depending on your organization's needs. Resource-based policies list the principals that are allowed or denied access to your resources, allowing you to set up policies such as cross-account access. Identity policies are specifically attached to users, groups, and roles within IAM.

For a step-by-step example, see Restrict access to your AWS Glue Data Catalog with resource-level IAM permissions and resource-based policies on the AWS Big Data Blog.

The fine-grained access portion of the policy is defined within the `Resource` clause. This portion defines both the AWS Glue Data Catalog object that the action can be performed on, and what resulting objects get returned by that operation.

A *development endpoint* is an environment that you can use to develop and test your AWS Glue scripts. You can add, delete, or rotate the SSH key of a development endpoint.

As of September 4, 2018, AWS KMS (*bring your own key* and *server-side encryption*) for AWS Glue ETL and the AWS Glue Data Catalog is supported.

AWS Glue dependency on other AWS services

For a user to work with the AWS Glue console, that user must have a minimum set of permissions that allows them to work with the AWS Glue resources for their AWS account. In addition to these AWS Glue permissions, the console requires permissions from the following services:

- Amazon CloudWatch Logs permissions to display logs.
- AWS Identity and Access Management (IAM) permissions to list and pass roles.
- AWS CloudFormation permissions to work with stacks.
- Amazon Elastic Compute Cloud (Amazon EC2) permissions to list virtual private clouds (VPCs), subnets, security groups, instances, and other objects (to set up Amazon EC2 items such as VPCs when running jobs, crawlers, and creating development endpoints).
- Amazon Simple Storage Service (Amazon S3) permissions to list buckets and objects, and to retrieve and save scripts.
- Amazon Redshift permissions to work with clusters.
- Amazon Relational Database Service (Amazon RDS) permissions to list instances.

Development endpoints

A development endpoint is an environment that you can use to develop and test your AWS Glue scripts. You can use AWS Glue to create, edit, and delete development endpoints. You can list all the development endpoints that are created. You can add, delete, or rotate the SSH key of a development endpoint. You can also create notebooks that use the development endpoint.

You provide configuration values to provision the development environments. These values tell AWS Glue how to set up the network so that you can access the development endpoint securely, and so that your endpoint can access your data stores. Then, you can create a notebook that connects to the development endpoint. You use your notebook to author and test your ETL script.

Use an AWS Identity and Access Management (IAM) role with permissions similar to the IAM role that you use to run AWS Glue ETL jobs. Use a virtual private cloud (VPC), a subnet, and a security group to create a development endpoint that can connect to your data resources securely. You generate an SSH key pair to connect to the development environment using SSH.

You can create development endpoints for Amazon S3 data and within a VPC that you can use to access datasets using JDBC.

You can install a Jupyter notebook client on your local machine and use it to debug and test ETL scripts on a development endpoint. Or, you can use a Sagemaker notebook to author ETL scripts in JupyterLab on AWS. See Use a SageMaker notebook with your development endpoint.

AWS Glue tags Amazon EC2 instances with a name that is prefixed with `aws-glue-dev-endpoint`.

You can set up a notebook server on a development endpoint to run PySpark with AWS Glue extensions.

Identity and access management for AWS Glue

AWS Identity and Access Management (IAM) is an AWS service that helps an administrator securely control access to AWS resources. IAM administrators control who can be *authenticated* (signed in) and *authorized* (have permissions) to use AWS Glue resources. IAM is an AWS service that you can use with no additional charge.

> **Note**
>
> You can grant access to your data in the AWS Glue Data Catalog using either AWS Glue methods or AWS Lake Formation grants. You can use AWS Identity and Access Management (IAM) policies to set fine-grained access control with AWS Glue methods. Lake Formation uses a simpler `GRANT/REVOKE` permissions model that is similar to the `GRANT/REVOKE` commands in a relational database system.
> This section includes information about how to use the AWS Glue methods. For information about using Lake Formation grants, see Granting Lake Formation permissions in the *AWS Lake Formation Developer Guide*.

Topics

- Audience
- Authenticating with identities
- Managing access using policies
- How AWS Glue works with IAM
- Configuring IAM permissions for AWS Glue
- AWS Glue access control policy examples
- Granting AWS managed policies for AWS Glue
- Granting dynamically scoped policies for job execution
- Specifying AWS Glue resource ARNs
- Granting cross-account access
- Troubleshooting AWS Glue identity and access

Audience

How you use AWS Identity and Access Management (IAM) differs, depending on the work that you do in AWS Glue.

Service user – If you use the AWS Glue service to do your job, then your administrator provides you with the credentials and permissions that you need. As you use more AWS Glue features to do your work, you might need additional permissions. Understanding how access is managed can help you request the right permissions from your administrator. If you cannot access a feature in AWS Glue, see Troubleshooting AWS Glue identity and access.

Service administrator – If you're in charge of AWS Glue resources at your company, you probably have full access to AWS Glue. It's your job to determine which AWS Glue features and resources your service users should access. You must then submit requests to your IAM administrator to change the permissions of your service users. Review the information on this page to understand the basic concepts of IAM. To learn more about how your company can use IAM with AWS Glue, see How AWS Glue works with IAM.

IAM administrator – If you're an IAM administrator, you might want to learn details about how you can write policies to manage access to AWS Glue. To view example AWS Glue identity-based policies that you can use in IAM, see Identity-based policy examples for AWS Glue.

Authenticating with identities

Authentication is how you sign in to AWS using your identity credentials. You must be *authenticated* (signed in to AWS) as the AWS account root user, as an IAM user, or by assuming an IAM role.

You can sign in to AWS as a federated identity by using credentials provided through an identity source. AWS IAM Identity Center (IAM Identity Center) users, your company's single sign-on authentication, and your Google or Facebook credentials are examples of federated identities. When you sign in as a federated identity, your administrator previously set up identity federation using IAM roles. When you access AWS by using federation, you are indirectly assuming a role.

Depending on the type of user you are, you can sign in to the AWS Management Console or the AWS access portal. For more information about signing in to AWS, see How to sign in to your AWS account in the *AWS Sign-In User Guide*.

If you access AWS programmatically, AWS provides a software development kit (SDK) and a command line interface (CLI) to cryptographically sign your requests by using your credentials. If you don't use AWS tools, you must sign requests yourself. For more information about using the recommended method to sign requests yourself, see AWS Signature Version 4 for API requests in the *IAM User Guide*.

Regardless of the authentication method that you use, you might be required to provide additional security information. For example, AWS recommends that you use multi-factor authentication (MFA) to increase the security of your account. To learn more, see Multi-factor authentication in the *AWS IAM Identity Center User Guide* and AWS Multi-factor authentication in IAM in the *IAM User Guide*.

AWS account root user

When you create an AWS account, you begin with one sign-in identity that has complete access to all AWS services and resources in the account. This identity is called the AWS account *root user* and is accessed by signing in with the email address and password that you used to create the account. We strongly recommend that you don't use the root user for your everyday tasks. Safeguard your root user credentials and use them to perform the tasks that only the root user can perform. For the complete list of tasks that require you to sign in as the root user, see Tasks that require root user credentials in the *IAM User Guide*.

Federated identity

As a best practice, require human users, including users that require administrator access, to use federation with an identity provider to access AWS services by using temporary credentials.

A *federated identity* is a user from your enterprise user directory, a web identity provider, the AWS Directory Service, the Identity Center directory, or any user that accesses AWS services by using credentials provided through an identity source. When federated identities access AWS accounts, they assume roles, and the roles provide temporary credentials.

For centralized access management, we recommend that you use AWS IAM Identity Center. You can create users and groups in IAM Identity Center, or you can connect and synchronize to a set of users and groups in your own identity source for use across all your AWS accounts and applications. For information about IAM Identity Center, see What is IAM Identity Center? in the *AWS IAM Identity Center User Guide*.

IAM users and groups

An *IAM user* is an identity within your AWS account that has specific permissions for a single person or application. Where possible, we recommend relying on temporary credentials instead of creating

IAM users who have long-term credentials such as passwords and access keys. However, if you have specific use cases that require long-term credentials with IAM users, we recommend that you rotate access keys. For more information, see Rotate access keys regularly for use cases that require long-term credentials in the *IAM User Guide*.

An *IAM group* is an identity that specifies a collection of IAM users. You can't sign in as a group. You can use groups to specify permissions for multiple users at a time. Groups make permissions easier to manage for large sets of users. For example, you could have a group named *IAMAdmins* and give that group permissions to administer IAM resources.

Users are different from roles. A user is uniquely associated with one person or application, but a role is intended to be assumable by anyone who needs it. Users have permanent long-term credentials, but roles provide temporary credentials. To learn more, see Use cases for IAM users in the *IAM User Guide*.

IAM roles

An *IAM role* is an identity within your AWS account that has specific permissions. It is similar to an IAM user, but is not associated with a specific person. To temporarily assume an IAM role in the AWS Management Console, you can switch from a user to an IAM role (console). You can assume a role by calling an AWS CLI or AWS API operation or by using a custom URL. For more information about methods for using roles, see Methods to assume a role in the *IAM User Guide*.

IAM roles with temporary credentials are useful in the following situations:

- **Federated user access** – To assign permissions to a federated identity, you create a role and define permissions for the role. When a federated identity authenticates, the identity is associated with the role and is granted the permissions that are defined by the role. For information about roles for federation, see Create a role for a third-party identity provider (federation) in the *IAM User Guide*. If you use IAM Identity Center, you configure a permission set. To control what your identities can access after they authenticate, IAM Identity Center correlates the permission set to a role in IAM. For information about permissions sets, see Permission sets in the *AWS IAM Identity Center User Guide*.
- **Temporary IAM user permissions** – An IAM user or role can assume an IAM role to temporarily take on different permissions for a specific task.
- **Cross-account access** – You can use an IAM role to allow someone (a trusted principal) in a different account to access resources in your account. Roles are the primary way to grant cross-account access. However, with some AWS services, you can attach a policy directly to a resource

(instead of using a role as a proxy). To learn the difference between roles and resource-based policies for cross-account access, see Cross account resource access in IAM in the *IAM User Guide*.

- **Cross-service access** – Some AWS services use features in other AWS services. For example, when you make a call in a service, it's common for that service to run applications in Amazon EC2 or store objects in Amazon S3. A service might do this using the calling principal's permissions, using a service role, or using a service-linked role.
  - **Forward access sessions (FAS)** – When you use an IAM user or role to perform actions in AWS, you are considered a principal. When you use some services, you might perform an action that then initiates another action in a different service. FAS uses the permissions of the principal calling an AWS service, combined with the requesting AWS service to make requests to downstream services. FAS requests are only made when a service receives a request that requires interactions with other AWS services or resources to complete. In this case, you must have permissions to perform both actions. For policy details when making FAS requests, see Forward access sessions.
  - **Service role** – A service role is an IAM role that a service assumes to perform actions on your behalf. An IAM administrator can create, modify, and delete a service role from within IAM. For more information, see Create a role to delegate permissions to an AWS service in the *IAM User Guide*.
  - **Service-linked role** – A service-linked role is a type of service role that is linked to an AWS service. The service can assume the role to perform an action on your behalf. Service-linked roles appear in your AWS account and are owned by the service. An IAM administrator can view, but not edit the permissions for service-linked roles.
- **Applications running on Amazon EC2** – You can use an IAM role to manage temporary credentials for applications that are running on an EC2 instance and making AWS CLI or AWS API requests. This is preferable to storing access keys within the EC2 instance. To assign an AWS role to an EC2 instance and make it available to all of its applications, you create an instance profile that is attached to the instance. An instance profile contains the role and enables programs that are running on the EC2 instance to get temporary credentials. For more information, see Use an IAM role to grant permissions to applications running on Amazon EC2 instances in the *IAM User Guide*.

Managing access using policies

You control access in AWS by creating policies and attaching them to AWS identities or resources. A policy is an object in AWS that, when associated with an identity or resource, defines their permissions. AWS evaluates these policies when a principal (user, root user, or role session) makes a request. Permissions in the policies determine whether the request is allowed or denied. Most policies are stored in AWS as JSON documents. For more information about the structure and contents of JSON policy documents, see Overview of JSON policies in the *IAM User Guide*.

Administrators can use AWS JSON policies to specify who has access to what. That is, which **principal** can perform **actions** on what **resources**, and under what **conditions**.

By default, users and roles have no permissions. To grant users permission to perform actions on the resources that they need, an IAM administrator can create IAM policies. The administrator can then add the IAM policies to roles, and users can assume the roles.

IAM policies define permissions for an action regardless of the method that you use to perform the operation. For example, suppose that you have a policy that allows the `iam:GetRole` action. A user with that policy can get role information from the AWS Management Console, the AWS CLI, or the AWS API.

Identity-based policies

Identity-based policies are JSON permissions policy documents that you can attach to an identity, such as an IAM user, group of users, or role. These policies control what actions users and roles can perform, on which resources, and under what conditions. To learn how to create an identity-based policy, see Define custom IAM permissions with customer managed policies in the *IAM User Guide*.

Identity-based policies can be further categorized as *inline policies* or *managed policies*. Inline policies are embedded directly into a single user, group, or role. Managed policies are standalone policies that you can attach to multiple users, groups, and roles in your AWS account. Managed policies include AWS managed policies and customer managed policies. To learn how to choose between a managed policy or an inline policy, see Choose between managed policies and inline policies in the *IAM User Guide*.

Resource-based policies

Resource-based policies are JSON policy documents that you attach to a resource. Examples of resource-based policies are IAM *role trust policies* and Amazon S3 *bucket policies*. In services that support resource-based policies, service administrators can use them to control access to a specific resource. For the resource where the policy is attached, the policy defines what actions a specified principal can perform on that resource and under what conditions. You must specify a principal in a resource-based policy. Principals can include accounts, users, roles, federated users, or AWS services.

Resource-based policies are inline policies that are located in that service. You can't use AWS managed policies from IAM in a resource-based policy.

Access control lists (ACLs)

Access control lists (ACLs) control which principals (account members, users, or roles) have permissions to access a resource. ACLs are similar to resource-based policies, although they do not use the JSON policy document format.

Amazon S3, AWS WAF, and Amazon VPC are examples of services that support ACLs. To learn more about ACLs, see Access control list (ACL) overview in the *Amazon Simple Storage Service Developer Guide*.

Other policy types

AWS supports additional, less-common policy types. These policy types can set the maximum permissions granted to you by the more common policy types.

- **Permissions boundaries** – A permissions boundary is an advanced feature in which you set the maximum permissions that an identity-based policy can grant to an IAM entity (IAM user or role). You can set a permissions boundary for an entity. The resulting permissions are the intersection of an entity's identity-based policies and its permissions boundaries. Resource-based policies that specify the user or role in the `Principal` field are not limited by the permissions boundary. An explicit deny in any of these policies overrides the allow. For more information about permissions boundaries, see Permissions boundaries for IAM entities in the *IAM User Guide*.
- **Service control policies (SCPs)** – SCPs are JSON policies that specify the maximum permissions for an organization or organizational unit (OU) in AWS Organizations. AWS Organizations is a service for grouping and centrally managing multiple AWS accounts that your business owns. If you enable all features in an organization, then you can apply service control policies (SCPs) to any or all of your accounts. The SCP limits permissions for entities in member accounts, including each AWS account root user. For more information about Organizations and SCPs, see Service control policies in the *AWS Organizations User Guide*.
- **Resource control policies (RCPs)** – RCPs are JSON policies that you can use to set the maximum available permissions for resources in your accounts without updating the IAM policies attached to each resource that you own. The RCP limits permissions for resources in member accounts and can impact the effective permissions for identities, including the AWS account root user, regardless of whether they belong to your organization. For more information about

Organizations and RCPs, including a list of AWS services that support RCPs, see Resource control policies (RCPs) in the *AWS Organizations User Guide*.

- **Session policies –** Session policies are advanced policies that you pass as a parameter when you programmatically create a temporary session for a role or federated user. The resulting session's permissions are the intersection of the user or role's identity-based policies and the session policies. Permissions can also come from a resource-based policy. An explicit deny in any of these policies overrides the allow. For more information, see Session policies in the *IAM User Guide*.

Multiple policy types

When multiple types of policies apply to a request, the resulting permissions are more complicated to understand. To learn how AWS determines whether to allow a request when multiple policy types are involved, see Policy evaluation logic in the *IAM User Guide*.

How AWS Glue works with IAM

Before you use IAM to manage access to AWS Glue, learn what IAM features are available to use with AWS Glue.

IAM features you can use with AWS Glue

IAM feature	AWS Glue support
Identity-based policies	Yes
Resource-based policies	Partial
Policy actions	Yes
Policy resources	Yes
Policy condition keys (service-specific)	Yes
ACLs	No
ABAC (tags in policies)	Partial
Temporary credentials	Yes

IAM feature	AWS Glue support
Principal permissions	No
Service roles	Yes
Service-linked roles	No

To get a high-level view of how AWS Glue and other AWS services work with most IAM features, see AWS services that work with IAM in the *IAM User Guide*.

Identity-based policies for AWS Glue

Supports identity-based policies: Yes

Identity-based policies are JSON permissions policy documents that you can attach to an identity, such as an IAM user, group of users, or role. These policies control what actions users and roles can perform, on which resources, and under what conditions. To learn how to create an identity-based policy, see Define custom IAM permissions with customer managed policies in the *IAM User Guide*.

With IAM identity-based policies, you can specify allowed or denied actions and resources as well as the conditions under which actions are allowed or denied. You can't specify the principal in an identity-based policy because it applies to the user or role to which it is attached. To learn about all of the elements that you can use in a JSON policy, see IAM JSON policy elements reference in the *IAM User Guide*.

AWS Glue supports identity-based policies (IAM policies) for all AWS Glue operations. By attaching a policy, you can grant permissions to create, access, or modify an AWS Glue resource, such as a table in the AWS Glue Data Catalog.

Identity-based policy examples for AWS Glue

To view examples of AWS Glue identity-based policies, see Identity-based policy examples for AWS Glue.

Resource-based policies within AWS Glue

Supports resource-based policies: Partial

Resource-based policies are JSON policy documents that you attach to a resource. Examples of resource-based policies are IAM *role trust policies* and Amazon S3 *bucket policies*. In services that support resource-based policies, service administrators can use them to control access to a specific resource. For the resource where the policy is attached, the policy defines what actions a specified principal can perform on that resource and under what conditions. You must specify a principal in a resource-based policy. Principals can include accounts, users, roles, federated users, or AWS services.

To enable cross-account access, you can specify an entire account or IAM entities in another account as the principal in a resource-based policy. Adding a cross-account principal to a resource-based policy is only half of establishing the trust relationship. When the principal and the resource are in different AWS accounts, an IAM administrator in the trusted account must also grant the principal entity (user or role) permission to access the resource. They grant permission by attaching an identity-based policy to the entity. However, if a resource-based policy grants access to a principal in the same account, no additional identity-based policy is required. For more information, see Cross account resource access in IAM in the *IAM User Guide*.

> **Note**
>
> You can only use an AWS Glue resource policy to manage permissions for Data Catalog resources. You can't attach it to any other AWS Glue resources such as jobs, triggers, development endpoints, crawlers, or classifiers.
> Only *one* resource policy is allowed per catalog, and its size is limited to 10 KB.

In AWS Glue, a resource policy is attached to a *catalog*, which is a virtual container for all the kinds of Data Catalog resources mentioned previously. Each AWS account owns a single catalog in an AWS Region whose catalog ID is the same as the AWS account ID. You cannot delete or modify a catalog.

A resource policy is evaluated for all API calls to the catalog where the caller principal is included in the `"Principal"` block of the policy document.

To view examples of AWS Glue resource-based policies, see Resource-based policy examples for AWS Glue.

Policy actions for AWS Glue

Supports policy actions: Yes

Administrators can use AWS JSON policies to specify who has access to what. That is, which **principal** can perform **actions** on what **resources**, and under what **conditions**.

The `Action` element of a JSON policy describes the actions that you can use to allow or deny access in a policy. Policy actions usually have the same name as the associated AWS API operation. There are some exceptions, such as *permission-only actions* that don't have a matching API operation. There are also some operations that require multiple actions in a policy. These additional actions are called *dependent actions*.

Include actions in a policy to grant permissions to perform the associated operation.

To see a list of AWS Glue actions, see Actions defined by AWS Glue in the *Service Authorization Reference*.

Policy actions in AWS Glue use the following prefix before the action:

```
glue
```

To specify multiple actions in a single statement, separate them with commas.

```
"Action": [
      "glue:action1",
      "glue:action2"
         ]
```

You can specify multiple actions using wildcards (*). For example, to specify all actions that begin with the word `Get`, include the following action:

```
"Action": "glue:Get*"
```

To view example policies, see AWS Glue access control policy examples.

Policy resources for AWS Glue

Supports policy resources: Yes

Administrators can use AWS JSON policies to specify who has access to what. That is, which **principal** can perform **actions** on what **resources**, and under what **conditions**.

The `Resource` JSON policy element specifies the object or objects to which the action applies. Statements must include either a `Resource` or a `NotResource` element. As a best practice, specify a resource using its Amazon Resource Name (ARN). You can do this for actions that support a specific resource type, known as *resource-level permissions*.

For actions that don't support resource-level permissions, such as listing operations, use a wildcard (*) to indicate that the statement applies to all resources.

```
"Resource": "*"
```

For more information about how to control access to AWS Glue resources using ARNs, see Specifying AWS Glue resource ARNs.

To see a list of AWS Glue resource types and their ARNs, see Resources defined by AWS Glue in the *Service Authorization Reference*. To learn which actions you can use to specify the ARN of each resource, see Actions defined by AWS Glue.

Policy condition keys for AWS Glue

Supports service-specific policy condition keys: Yes

Administrators can use AWS JSON policies to specify who has access to what. That is, which **principal** can perform **actions** on what **resources**, and under what **conditions**.

The `Condition` element (or `Condition` *block*) lets you specify conditions in which a statement is in effect. The `Condition` element is optional. You can create conditional expressions that use condition operators, such as equals or less than, to match the condition in the policy with values in the request.

If you specify multiple `Condition` elements in a statement, or multiple keys in a single `Condition` element, AWS evaluates them using a logical `AND` operation. If you specify multiple values for a single condition key, AWS evaluates the condition using a logical `OR` operation. All of the conditions must be met before the statement's permissions are granted.

You can also use placeholder variables when you specify conditions. For example, you can grant an IAM user permission to access a resource only if it is tagged with their IAM user name. For more information, see IAM policy elements: variables and tags in the *IAM User Guide*.

AWS supports global condition keys and service-specific condition keys. To see all AWS global condition keys, see AWS global condition context keys in the *IAM User Guide*.

To see a list of AWS Glue condition keys, see Condition keys for AWS Glue in the *Service Authorization Reference*. To learn which actions and resources you can use a condition key with, see Actions defined by AWS Glue.

To view example policies, see Control settings using condition keys or context keys.

ACLs in AWS Glue

Supports ACLs: No

Access control lists (ACLs) control which principals (account members, users, or roles) have permissions to access a resource. ACLs are similar to resource-based policies, although they do not use the JSON policy document format.

ABAC with AWS Glue

Supports ABAC (tags in policies): Partial

Attribute-based access control (ABAC) is an authorization strategy that defines permissions based on attributes. In AWS, these attributes are called *tags*. You can attach tags to IAM entities (users or roles) and to many AWS resources. Tagging entities and resources is the first step of ABAC. Then you design ABAC policies to allow operations when the principal's tag matches the tag on the resource that they are trying to access.

ABAC is helpful in environments that are growing rapidly and helps with situations where policy management becomes cumbersome.

To control access based on tags, you provide tag information in the condition element of a policy using the `aws:ResourceTag/key-name`, `aws:RequestTag/key-name`, or `aws:TagKeys` condition keys.

If a service supports all three condition keys for every resource type, then the value is **Yes** for the service. If a service supports all three condition keys for only some resource types, then the value is **Partial**.

For more information about ABAC, see Define permissions with ABAC authorization in the *IAM User Guide*. To view a tutorial with steps for setting up ABAC, see Use attribute-based access control (ABAC) in the *IAM User Guide*.

Important

The condition context keys apply only to AWS Glue API actions on crawlers, jobs, triggers, and development endpoints. For more information about which API operations are affected, see Condition keys for AWS Glue.

The AWS Glue Data Catalog API operations don't currently support the `aws:referer` and `aws:UserAgent` global condition context keys.

To view an example identity-based policy for limiting access to a resource based on the tags on that resource, see Grant access using tags.

Using temporary credentials with AWS Glue

Supports temporary credentials: Yes

Some AWS services don't work when you sign in using temporary credentials. For additional information, including which AWS services work with temporary credentials, see AWS services that work with IAM in the *IAM User Guide*.

You are using temporary credentials if you sign in to the AWS Management Console using any method except a user name and password. For example, when you access AWS using your company's single sign-on (SSO) link, that process automatically creates temporary credentials. You also automatically create temporary credentials when you sign in to the console as a user and then switch roles. For more information about switching roles, see Switch from a user to an IAM role (console) in the *IAM User Guide*.

You can manually create temporary credentials using the AWS CLI or AWS API. You can then use those temporary credentials to access AWS. AWS recommends that you dynamically generate temporary credentials instead of using long-term access keys. For more information, see Temporary security credentials in IAM.

Cross-service principal permissions for AWS Glue

Supports forward access sessions (FAS): No

When you use an IAM user or role to perform actions in AWS, you are considered a principal. When you use some services, you might perform an action that then initiates another action in a different service. FAS uses the permissions of the principal calling an AWS service, combined with the requesting AWS service to make requests to downstream services. FAS requests are only made when a service receives a request that requires interactions with other AWS services or resources to complete. In this case, you must have permissions to perform both actions. For policy details when making FAS requests, see Forward access sessions.

Service roles for AWS Glue

Supports service roles: Yes

A service role is an IAM role that a service assumes to perform actions on your behalf. An IAM administrator can create, modify, and delete a service role from within IAM. For more information, see Create a role to delegate permissions to an AWS service in the *IAM User Guide*.

> **Warning**
>
> Changing the permissions for a service role might break AWS Glue functionality. Edit service roles only when AWS Glue provides guidance to do so.

For detailed instructions on creating a service role for AWS Glue, see Step 1: Create an IAM policy for the AWS Glue service and Step 2: Create an IAM role for AWS Glue.

Service-linked roles for AWS Glue

Supports service-linked roles: No

A service-linked role is a type of service role that is linked to an AWS service. The service can assume the role to perform an action on your behalf. Service-linked roles appear in your AWS account and are owned by the service. An IAM administrator can view, but not edit the permissions for service-linked roles.

For details about creating or managing service-linked roles, see AWS services that work with IAM. Find a service in the table that includes a `Yes` in the **Service-linked role** column. Choose the **Yes** link to view the service-linked role documentation for that service.

Configuring IAM permissions for AWS Glue

You use AWS Identity and Access Management (IAM) to define policies and roles that AWS Glue uses to access resources. The following steps lead you through various options for setting up the permissions for AWS Glue. Depending on your business needs, you might have to add or reduce access to your resources.

> **Note**
>
> To get started with basic IAM permissions for AWS Glue instead, see Setting up IAM permissions for AWS Glue.

1. Create an IAM policy for the AWS Glue service: Create a service policy that allows access to AWS Glue resources.
2. Create an IAM role for AWS Glue: Create an IAM role, and attach the AWS Glue service policy and a policy for your Amazon Simple Storage Service (Amazon S3) resources that are used by AWS Glue.
3. Attach a policy to users or groups that access AWS Glue: Attach policies to any users or groups that sign in to the AWS Glue console.
4. Create an IAM policy for notebooks: Create a notebook server policy to use in the creation of notebook servers on development endpoints.
5. Create an IAM role for notebooks: Create an IAM role and attach the notebook server policy.
6. Create an IAM policy for Amazon SageMaker AI notebooks: Create an IAM policy to use when creating Amazon SageMaker AI notebooks on development endpoints.
7. Create an IAM role for Amazon SageMaker AI notebooks: Create an IAM role and attach the policy to grant permissions when creating Amazon SageMaker AI notebooks on development endpoints.

Step 1: Create an IAM policy for the AWS Glue service

For any operation that accesses data on another AWS resource, such as accessing your objects in Amazon S3, AWS Glue needs permission to access the resource on your behalf. You provide those permissions by using AWS Identity and Access Management (IAM).

> **Note**
>
> You can skip this step if you use the AWS managed policy **`AWSGlueServiceRole`**.

In this step, you create a policy that is similar to `AWSGlueServiceRole`. You can find the most current version of `AWSGlueServiceRole` on the IAM console.

To create an IAM policy for AWS Glue

This policy grants permission for some Amazon S3 actions to manage resources in your account that are needed by AWS Glue when it assumes the role using this policy. Some of the resources that are specified in this policy refer to default names that are used by AWS Glue for Amazon S3 buckets, Amazon S3 ETL scripts, CloudWatch Logs, and Amazon EC2 resources. For simplicity, AWS Glue writes some Amazon S3 objects into buckets in your account prefixed with `aws-glue-*` by default.

1. Sign in to the AWS Management Console and open the IAM console at https://console.aws.amazon.com/iam/.
2. In the left navigation pane, choose **Policies**.
3. Choose **Create Policy**.
4. On the **Create Policy** screen, navigate to a tab to edit JSON. Create a policy document with the following JSON statements, and then choose **Review policy**.

   > **Note**
   >
   > Add any permissions needed for Amazon S3 resources. You might want to scope the resources section of your access policy to only those resources that are required.

   JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:*",
        "s3:GetBucketLocation",
        "s3:ListBucket",
        "s3:ListAllMyBuckets",
        "s3:GetBucketAcl",
        "ec2:DescribeVpcEndpoints",
        "ec2:DescribeRouteTables",
        "ec2:CreateNetworkInterface",
        "ec2:DeleteNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DescribeSecurityGroups",
        "ec2:DescribeSubnets",
        "ec2:DescribeVpcAttribute",
        "iam:ListRolePolicies",
        "iam:GetRole",
        "iam:GetRolePolicy",
        "cloudwatch:PutMetricData"
      ],
      "Resource": [
        "*"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "s3:CreateBucket",
        "s3:PutBucketPublicAccessBlock"
      ],
      "Resource": [
        "arn:aws:s3:::aws-glue-*"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "s3:GetObject",
        "s3:PutObject",
        "s3:DeleteObject"
      ],
      "Resource": [
        "arn:aws:s3:::aws-glue-*/*",
        "arn:aws:s3:::*/*aws-glue-*/*"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "s3:GetObject"
      ],
      "Resource": [
        "arn:aws:s3:::crawler-public*",
        "arn:aws:s3:::aws-glue-*"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "logs:CreateLogGroup",
        "logs:CreateLogStream",
        "logs:PutLogEvents",
        "logs:AssociateKmsKey"
      ],
      "Resource": [
        "arn:aws:logs:*:*:log-group:/aws-glue/*"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "ec2:CreateTags",
        "ec2:DeleteTags"
      ],
      "Condition": {
        "ForAllValues:StringEquals": {
          "aws:TagKeys": [
            "aws-glue-service-resource"
          ]
        }
      },
      "Resource": [
        "arn:aws:ec2:*:*:network-interface/*",
        "arn:aws:ec2:*:*:security-group/*",
        "arn:aws:ec2:*:*:instance/*"
      ]
    }
  ]
}
```

The following table describes the permissions granted by this policy.

Action	Resource	Description
`"glue:*"`	`"*"`	Grants permission to run all AWS Glue API operations.

Action	Resource	Description
`"s3:GetBucketLocation", "s3:ListBucket", "s3:ListAllMyBuckets", "s3:GetBucketAcl",`	`"*"`	Allows listing of Amazon S3 buckets from crawlers, jobs, development endpoints, and notebook servers.
`"ec2:DescribeVpcEndpoints", "ec2:DescribeRouteTables", "ec2:CreateNetworkInterface", "ec2:DeleteNetworkInterface", "ec2:DescribeNetworkInterfaces", "ec2:DescribeSecurityGroups", "ec2:DescribeSubnets", "ec2:DescribeVpcAttribute",`	`"*"`	Allows the setup of Amazon EC2 network items, such as virtual private clouds (VPCs) when running jobs, crawlers, and development endpoints.
`"iam:ListRolePolicies", "iam:GetRole", "iam:GetRolePolicy"`	`"*"`	Allows listing IAM roles from crawlers, jobs, development endpoints, and notebook servers.
`"cloudwatch:PutMetricData"`	`"*"`	Allows writing CloudWatch metrics for jobs.

Action	Resource	Description
`"s3:CreateBucket", "s3:PutBucketPublicAccessBlock"`	`"arn:aws:s3:::aws-glue-*"`	Allows the creation of Amazon S3 buckets in your account from jobs and notebook servers. Naming convention: Uses Amazon S3 folders named **aws-glue-**. Enables AWS Glue to create buckets that block public access.
`"s3:GetObject", "s3:PutObject", "s3:DeleteObject"`	`"arn:aws:s3:::aws-glue-*/*", "arn:aws:s3:::*/*aws-glue-*/*"`	Allows get, put, and delete of Amazon S3 objects into your account when storing objects such as ETL scripts and notebook server locations. Naming convention: Grants permission to Amazon S3 buckets or folders whose names are prefixed with **aws-glue-**.
`"s3:GetObject"`	`"arn:aws:s3:::crawler-public*", "arn:aws:s3:::aws-glue-*"`	Allows get of Amazon S3 objects used by examples and tutorials from crawlers and jobs. Naming convention: Amazon S3 bucket names begin with **crawler-public** and **aws-glue-**.

Action	Resource	Description
`"logs:CreateLogGroup", "logs:CreateLogStream", "logs:PutLogEvents"`	`"arn:aws:logs:*:*:log-group:/aws-glue/*"`	Allows writing logs to CloudWatch Logs. Naming convention: AWS Glue writes logs to log groups whose names begin with **aws-glue**.
`"ec2:CreateTags", "ec2:DeleteTags"`	`"arn:aws:ec2:*:*:network-interface/*", "arn:aws:ec2:*:*:security-group/*", "arn:aws:ec2:*:*:instance/*"`	Allows tagging of Amazon EC2 resources created for development endpoints. Naming convention: AWS Glue tags Amazon EC2 network interfaces, security groups, and instances with **aws-glue-service-resource**.

5. On the **Review Policy** screen, enter your **Policy Name**, for example **GlueServiceRolePolicy**. Enter an optional description, and when you're satisfied with the policy, choose **Create policy**.

Step 2: Create an IAM role for AWS Glue

You need to grant your IAM role permissions that AWS Glue can assume when calling other services on your behalf. This includes access to Amazon S3 for any sources, targets, scripts, and temporary directories that you use with AWS Glue. Permission is needed by crawlers, jobs, and development endpoints.

You provide those permissions by using AWS Identity and Access Management (IAM). Add a policy to the IAM role that you pass to AWS Glue.

To create an IAM role for AWS Glue

1. Sign in to the AWS Management Console and open the IAM console at https://console.aws.amazon.com/iam/.
2. In the left navigation pane, choose **Roles**.
3. Choose **Create role**.

4. Choose **AWS service** as the trusted entity type. Then, for service or use case, find and choose **AWS Glue**. Choose **Next**.
5. On the **Add permissions** page, choose the policies that contain the required permissions; for example, the AWS managed policy `AWSGlueServiceRole` for general AWS Glue permissions and the AWS managed policy **AmazonS3FullAccess** for access to Amazon S3 resources. Then choose **Next**.

> **Note**
>
> Ensure that one of the policies in this role grants permissions to your Amazon S3 sources and targets. You might want to provide your own policy for access to specific Amazon S3 resources. Data sources require `s3:ListBucket` and `s3:GetObject` permissions. Data targets require `s3:ListBucket`, `s3:PutObject`, and `s3:DeleteObject` permissions. For more information about creating an Amazon S3 policy for your resources, see Specifying Resources in a Policy. For an example Amazon S3 policy, see Writing IAM Policies: How to Grant Access to an Amazon S3 Bucket.
>
> If you plan to access Amazon S3 sources and targets that are encrypted with SSE-KMS, attach a policy that allows AWS Glue crawlers, jobs, and development endpoints to decrypt the data. For more information, see Protecting Data Using Server-Side Encryption with AWS KMS-Managed Keys (SSE-KMS).
>
> The following is an example.
>
> JSON
>
> ```
> {
>   "Version": "2012-10-17",
>   "Statement": [
>     {
>       "Effect": "Allow",
>       "Action": [
>         "kms:Decrypt"
>       ],
>       "Resource": [
>         "arn:aws:kms:*:111122223333:key/key-id"
>       ]
>     }
>   ]
> ```

```
}
```

6. Name your role and add a description (optional), then review the trust policy and permissions. For **Role name**, enter a name for your role; for example, `AWSGlueServiceRoleDefault`. Create the role with the name prefixed with the string `AWSGlueServiceRole` to allow the role to be passed from console users to the service. AWS Glue provided policies expect IAM service roles to begin with `AWSGlueServiceRole`. Otherwise, you must add a policy to allow your users the `iam:PassRole` permission for IAM roles to match your naming convention. Choose **Create Role**.

> **Note**
>
> When you create a notebook with a role, that role is then passed to interactive sessions so that the same role can be used in both places. As such, the `iam:PassRole` permission needs to be part of the role's policy.
> Create a new policy for your role using the following example. Replace the account number with your own and the role name.
>
> JSON
>
> ```
> {
>   "Version": "2012-10-17",
>   "Statement": [
>     {
>       "Effect": "Allow",
>       "Action": "iam:PassRole",
>       "Resource": "arn:aws:iam::090000000210:role/<role_name>"
>     }
>   ]
> }
> ```

7. Add tags to your role (optional). Tags are key-value pairs that you can add to AWS resources to help identify, organize, or search for resources. Then, choose **Create role**.

Step 3: Attach a policy to users or groups that access AWS Glue

The administrator must assign permissions to any users, groups, or roles using the AWS Glue console or AWS Command Line Interface (AWS CLI). You provide those permissions by using AWS Identity and Access Management (IAM), through policies. This step describes assigning permissions to users or groups.

When you finish this step, your user or group has the following policies attached:

- The AWS managed policy `AWSGlueConsoleFullAccess` or the custom policy **GlueConsoleAccessPolicy**
- **`AWSGlueConsoleSageMakerNotebookFullAccess`**
- **`CloudWatchLogsReadOnlyAccess`**
- **`AWSCloudFormationReadOnlyAccess`**
- **`AmazonAthenaFullAccess`**

To attach an inline policy and embed it in a user or group

You can attach an AWS managed policy or an inline policy to a user or group to access the AWS Glue console. Some of the resources specified in this policy refer to default names that are used by AWS Glue for Amazon S3 buckets, Amazon S3 ETL scripts, CloudWatch Logs, AWS CloudFormation, and Amazon EC2 resources. For simplicity, AWS Glue writes some Amazon S3 objects into buckets in your account prefixed with `aws-glue-*` by default.

> **Note**
>
> You can skip this step if you use the AWS managed policy **`AWSGlueConsoleFullAccess`**.

> **Important**
>
> AWS Glue needs permission to assume a role that is used to perform work on your behalf. **To accomplish this, you add the `iam:PassRole` permissions to your AWS Glue users or groups.** This policy grants permission to roles that begin with `AWSGlueServiceRole` for AWS Glue service roles, and `AWSGlueServiceNotebookRole` for roles that are required when you create a notebook server. You can also create your own policy for `iam:PassRole` permissions that follows your naming convention.

Per security best practices, it is recommended to restrict access by tightening policies to further restrict access to Amazon S3 bucket and Amazon CloudWatch log groups. For an example Amazon S3 policy, see [Writing IAM Policies: How to Grant Access to an Amazon S3 Bucket](https://aws.amazon.com/blogs/security/writing-iam-policies-how-to-grant-access-to-an-amazon-s3-bucket/).

In this step, you create a policy that is similar to `AWSGlueConsoleFullAccess`. You can find the most current version of `AWSGlueConsoleFullAccess` on the IAM console.

1. Sign in to the AWS Management Console and open the IAM console at https://console.aws.amazon.com/iam/.
2. In the navigation pane, choose **Users** or **User groups**.
3. In the list, choose the name of the user or group to embed a policy in.
4. Choose the **Permissions** tab and, if necessary, expand the **Permissions policies** section.
5. Choose the **Add Inline policy** link.
6. On the **Create Policy** screen, navigate to a tab to edit JSON. Create a policy document with the following JSON statements, and then choose **Review policy**.

   JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:*",
        "redshift:DescribeClusters",
        "redshift:DescribeClusterSubnetGroups",
        "iam:ListRoles",
        "iam:ListUsers",
        "iam:ListGroups",
        "iam:ListRolePolicies",
        "iam:GetRole",
        "iam:GetRolePolicy",
        "iam:ListAttachedRolePolicies",
        "ec2:DescribeSecurityGroups",
        "ec2:DescribeSubnets",
        "ec2:DescribeVpcs",
        "ec2:DescribeVpcEndpoints",
        "ec2:DescribeRouteTables",
        "ec2:DescribeVpcAttribute",
        "ec2:DescribeKeyPairs",
        "ec2:DescribeInstances",
        "rds:DescribeDBInstances",
        "rds:DescribeDBClusters",
        "rds:DescribeDBSubnetGroups",
        "s3:ListAllMyBuckets",
        "s3:ListBucket",
        "s3:GetBucketAcl",
        "s3:GetBucketLocation",
        "cloudformation:DescribeStacks",
        "cloudformation:GetTemplateSummary",
        "dynamodb:ListTables",
        "kms:ListAliases",
        "kms:DescribeKey",
        "cloudwatch:GetMetricData",
        "cloudwatch:ListDashboards"
      ],
      "Resource": [
        "*"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "s3:GetObject",
        "s3:PutObject"
      ],
      "Resource": [
        "arn:aws:s3:::*/*aws-glue-*/*",
        "arn:aws:s3:::aws-glue-*"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "tag:GetResources"
      ],
      "Resource": [
        "*"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "s3:CreateBucket",
        "s3:PutBucketPublicAccessBlock"
      ],
      "Resource": [
        "arn:aws:s3:::aws-glue-*"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "logs:GetLogEvents"
      ],
      "Resource": [
        "arn:aws:logs:*:*:/aws-glue/*"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "cloudformation:CreateStack",
        "cloudformation:DeleteStack"
      ],
      "Resource": "arn:aws:cloudformation:*:*:stack/aws-glue*/*"
    },
    {
      "Effect": "Allow",
      "Action": [
        "ec2:RunInstances"
      ],
      "Resource": [
        "arn:aws:ec2:*:*:instance/*",
        "arn:aws:ec2:*:*:key-pair/*",
        "arn:aws:ec2:*:*:image/*",
        "arn:aws:ec2:*:*:security-group/*",
        "arn:aws:ec2:*:*:network-interface/*",
        "arn:aws:ec2:*:*:subnet/*",
        "arn:aws:ec2:*:*:volume/*"
      ]
    },
    {
      "Action": [
        "iam:PassRole"
      ],
      "Effect": "Allow",
      "Resource": "arn:aws:iam::*:role/AWSGlueServiceRole*",
      "Condition": {
        "StringLike": {
          "iam:PassedToService": [
            "glue.amazonaws.com"
          ]
        }
      }
    },
    {
      "Action": [
        "iam:PassRole"
      ],
      "Effect": "Allow",
      "Resource": "arn:aws:iam::*:role/AWSGlueServiceNotebookRole*",
      "Condition": {
        "StringLike": {
          "iam:PassedToService": [
            "ec2.amazonaws.com"
          ]
        }
      }
    },
    {
      "Action": [
        "iam:PassRole"
      ],
      "Effect": "Allow",
      "Resource": [
        "arn:aws:iam::*:role/service-role/AWSGlueServiceRole*"
      ],
      "Condition": {
        "StringLike": {
          "iam:PassedToService": [
            "glue.amazonaws.com"
          ]
        }
      }
    }
  ]
```

```
}
```

The following table describes the permissions granted by this policy.

Action	Resource	Description
`"glue:*"`	`"*"`	Grants permission to run all AWS Glue API operations. If you had previously created your policy without the `"glue:*"` action, you must add the following individual permissions to your policy: • "glue:ListCrawlers" • "glue:BatchGetCrawlers" • "glue:ListTriggers" • "glue:BatchGetTriggers" • "glue:ListDevEndpoints" • "glue:BatchGetDevEndpoints" • "glue:ListJobs" • "glue:BatchGetJobs"
`"redshift:DescribeClusters", "redshift:DescribeClusterSubnetGroups"`	`"*"`	Allows creation of connections to Amazon Redshift.

Action	Resource	Description
`"iam:ListRoles", "iam:ListRolePolicies", "iam:GetRole", "iam:GetRolePolicy", "iam:ListAttachedRolePolicies"`	`"*"`	Allows listing IAM roles when working with crawlers, jobs, development endpoints, and notebook servers.
`"ec2:DescribeSecurityGroups", "ec2:DescribeSubnets", "ec2:DescribeVpcs", "ec2:DescribeVpcEndpoints", "ec2:DescribeRouteTables", "ec2:DescribeVpcAttribute", "ec2:DescribeKeyPairs", "ec2:DescribeInstances"`	`"*"`	Allows setup of Amazon EC2 network items, such as VPCs, when running jobs, crawlers, and development endpoints.
`"rds:DescribeDBInstances"`	`"*"`	Allows creation of connections to Amazon RDS.
`"s3:ListAllMyBuckets", "s3:ListBucket", "s3:GetBucketAcl", "s3:GetBucketLocation"`	`"*"`	Allows listing of Amazon S3 buckets when working with crawlers, jobs, development endpoints, and notebook servers.
`"dynamodb:ListTables"`	`"*"`	Allows listing of DynamoDB tables.
`"kms:ListAliases", "kms:DescribeKey"`	`"*"`	Allows working with KMS keys.

Action	Resource	Description
`"cloudwatch:GetMetricData", "cloudwatch:ListDashboards"`	`"*"`	Allows working with CloudWatch metrics.
`"s3:GetObject", "s3:PutObject"`	`"arn:aws:s3::: aws-glue-*/*", "arn:aws:s3::: */*aws-glue-*/*", "arn:aws:s3::: aws-glue-*"`	Allows get and put of Amazon S3 objects into your account when storing objects such as ETL scripts and notebook server locations. Naming convention: Grants permission to Amazon S3 buckets or folders whose names are prefixed with **aws-glue-**.
`"tag:GetResources"`	`"*"`	Allows retrieval of AWS tags.

Action	Resource	Description
`"s3:CreateBucket", "s3:PutBucketPublicAccessBlock"`	`"arn:aws:s3::: aws-glue-*"`	Allows creation of an Amazon S3 bucket into your account when storing objects such as ETL scripts and notebook server locations. Naming convention: Grants permission to Amazon S3 buckets whose names are prefixed with **aws-glue-**. Enables AWS Glue to create buckets that block public access.
`"logs:GetLogEvents"`	`"arn:aws:logs:*:*: /aws-glue/*"`	Allows retrieval of CloudWatch Logs. Naming convention: AWS Glue writes logs to log groups whose names begin with **aws-glue-**.

Action	Resource	Description
`"cloudformation:CreateStack", "cloudformation:DeleteStack"`	`"arn:aws:cloudformation:*:*:stack/aws-glue*/*"`	Allows managing AWS CloudFormation stacks when working with notebook servers. Naming convention: AWS Glue creates stacks whose names begin with **aws-glue**.
`"ec2:RunInstances"`	`"arn:aws:ec2:*:*:instance/*", "arn:aws:ec2:*:*:key-pair/*", "arn:aws:ec2:*:*:image/*", "arn:aws:ec2:*:*:security-group/*", "arn:aws:ec2:*:*:network-interface/*", "arn:aws:ec2:*:*:subnet/*", "arn:aws:ec2:*:*:volume/*"`	Allows running of development endpoints and notebook servers.
`"iam:PassRole"`	`"arn:aws:iam::*:role/ AWSGlueServiceRole*"`	Allows AWS Glue to assume `PassRole` permission for roles that begin with `AWSGlueServiceRole` .

Action	Resource	Description
`"iam:PassRole"`	`"arn:aws:iam::*:role/ AWSGlueServiceNotebookRole*"`	Allows Amazon EC2 to assume `PassRole` permission for roles that begin with `AWSGlueServiceNotebookRole` .
`"iam:PassRole"`	`"arn:aws:iam::*:role/service-role/AWSGlueServiceRole*"`	Allows AWS Glue to assume `PassRole` permission for roles that begin with `service-role/AWSGlueServiceRole` .

7. On the **Review policy** screen, enter a name for the policy, for example **GlueConsoleAccessPolicy**. When you're satisfied with the policy, choose **Create policy**. Ensure that no errors appear in a red box at the top of the screen. Correct any that are reported.

> **Note**
>
> If **Use autoformatting** is selected, the policy is reformatted whenever you open a policy or choose **Validate Policy**.

To attach the AWSGlueConsoleFullAccess managed policy

You can attach the `AWSGlueConsoleFullAccess` policy to provide permissions that are required by the AWS Glue console user.

> **Note**
>
> You can skip this step if you created your own policy for AWS Glue console access.

1. Sign in to the AWS Management Console and open the IAM console at https://console.aws.amazon.com/iam/.
2. In the navigation pane, choose **Policies**.
3. In the list of policies, select the check box next to the **AWSGlueConsoleFullAccess**. You can use the **Filter** menu and the search box to filter the list of policies.
4. Choose **Policy actions**, and then choose **Attach**.
5. Choose the user to attach the policy to. You can use the **Filter** menu and the search box to filter the list of principal entities. After choosing the user to attach the policy to, choose **Attach policy**.

To attach the `AWSGlueConsoleSageMakerNotebookFullAccess` managed policy

You can attach the `AWSGlueConsoleSageMakerNotebookFullAccess` policy to a user to manage SageMaker AI notebooks created on the AWS Glue console. In addition to other required AWS Glue console permissions, this policy grants access to resources needed to manage SageMaker AI notebooks.

1. Sign in to the AWS Management Console and open the IAM console at https://console.aws.amazon.com/iam/.
2. In the navigation pane, choose **Policies**.
3. In the list of policies, select the check box next to the **AWSGlueConsoleSageMakerNotebookFullAccess**. You can use the **Filter** menu and the search box to filter the list of policies.
4. Choose **Policy actions**, and then choose **Attach**.
5. Choose the user to attach the policy to. You can use the **Filter** menu and the search box to filter the list of principal entities. After choosing the user to attach the policy to, choose **Attach policy**.

To attach the CloudWatchLogsReadOnlyAccess managed policy

You can attach the **CloudWatchLogsReadOnlyAccess** policy to a user to view the logs created by AWS Glue on the CloudWatch Logs console.

1. Sign in to the AWS Management Console and open the IAM console at https://console.aws.amazon.com/iam/.

2. In the navigation pane, choose **Policies**.
3. In the list of policies, select the check box next to the **CloudWatchLogsReadOnlyAccess**. You can use the **Filter** menu and the search box to filter the list of policies.
4. Choose **Policy actions**, and then choose **Attach**.
5. Choose the user to attach the policy to. You can use the **Filter** menu and the search box to filter the list of principal entities. After choosing the user to attach the policy to, choose **Attach policy**.

To attach the AWSCloudFormationReadOnlyAccess managed policy

You can attach the **AWSCloudFormationReadOnlyAccess** policy to a user to view the AWS CloudFormation stacks used by AWS Glue on the AWS CloudFormation console.

1. Sign in to the AWS Management Console and open the IAM console at https://console.aws.amazon.com/iam/.
2. In the navigation pane, choose **Policies**.
3. In the list of policies, select the check box next to **AWSCloudFormationReadOnlyAccess**. You can use the **Filter** menu and the search box to filter the list of policies.
4. Choose **Policy actions**, and then choose **Attach**.
5. Choose the user to attach the policy to. You can use the **Filter** menu and the search box to filter the list of principal entities. After choosing the user to attach the policy to, choose **Attach policy**.

To attach the AmazonAthenaFullAccess managed policy

You can attach the **AmazonAthenaFullAccess** policy to a user to view Amazon S3 data in the Athena console.

1. Sign in to the AWS Management Console and open the IAM console at https://console.aws.amazon.com/iam/.
2. In the navigation pane, choose **Policies**.
3. In the list of policies, select the check box next to the **AmazonAthenaFullAccess**. You can use the **Filter** menu and the search box to filter the list of policies.
4. Choose **Policy actions**, and then choose **Attach**.

5. Choose the user to attach the policy to. You can use the **Filter** menu and the search box to filter the list of principal entities. After choosing the user to attach the policy to, choose **Attach policy**.

Step 4: Create an IAM policy for notebook servers

If you plan to use notebooks with development endpoints, you must specify permissions when you create the notebook server. You provide those permissions by using AWS Identity and Access Management (IAM).

This policy grants permission for some Amazon S3 actions to manage resources in your account that are needed by AWS Glue when it assumes the role using this policy. Some of the resources that are specified in this policy refer to default names used by AWS Glue for Amazon S3 buckets, Amazon S3 ETL scripts, and Amazon EC2 resources. For simplicity, AWS Glue defaults writing some Amazon S3 objects into buckets in your account prefixed with `aws-glue-*`.

> **Note**
>
> You can skip this step if you use the AWS managed policy **`AWSGlueServiceNotebookRole`**.

In this step, you create a policy that is similar to `AWSGlueServiceNotebookRole`. You can find the most current version of `AWSGlueServiceNotebookRole` on the IAM console.

To create an IAM policy for notebooks

1. Sign in to the AWS Management Console and open the IAM console at https://console.aws.amazon.com/iam/.
2. In the left navigation pane, choose **Policies**.
3. Choose **Create Policy**.
4. On the **Create Policy** screen, navigate to a tab to edit JSON. Create a policy document with the following JSON statements, and then choose **Review policy**.

   JSON

   ```
   {
     "Version": "2012-10-17",
   ```

```
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:CreateDatabase",
        "glue:CreatePartition",
        "glue:CreateTable",
        "glue:DeleteDatabase",
        "glue:DeletePartition",
        "glue:DeleteTable",
        "glue:GetDatabase",
        "glue:GetDatabases",
        "glue:GetPartition",
        "glue:GetPartitions",
        "glue:GetTable",
        "glue:GetTableVersions",
        "glue:GetTables",
        "glue:UpdateDatabase",
        "glue:UpdatePartition",
        "glue:UpdateTable",
        "glue:GetJobBookmark",
        "glue:ResetJobBookmark",
        "glue:CreateConnection",
        "glue:CreateJob",
        "glue:DeleteConnection",
        "glue:DeleteJob",
        "glue:GetConnection",
        "glue:GetConnections",
        "glue:GetDevEndpoint",
        "glue:GetDevEndpoints",
        "glue:GetJob",
        "glue:GetJobs",
        "glue:UpdateJob",
        "glue:BatchDeleteConnection",
        "glue:UpdateConnection",
        "glue:GetUserDefinedFunction",
        "glue:UpdateUserDefinedFunction",
        "glue:GetUserDefinedFunctions",
        "glue:DeleteUserDefinedFunction",
        "glue:CreateUserDefinedFunction",
        "glue:BatchGetPartition",
        "glue:BatchDeletePartition",
        "glue:BatchCreatePartition",
        "glue:BatchDeleteTable",
          "glue:UpdateDevEndpoint",
          "s3:GetBucketLocation",
          "s3:ListBucket",
          "s3:ListAllMyBuckets",
          "s3:GetBucketAcl"
        ],
        "Resource": [
          "*"
        ]
      },
      {
        "Effect": "Allow",
        "Action": [
          "s3:GetObject"
        ],
        "Resource": [
          "arn:aws:s3:::crawler-public*",
          "arn:aws:s3:::aws-glue*"
        ]
      },
      {
        "Effect": "Allow",
        "Action": [
          "s3:PutObject",
          "s3:DeleteObject"
        ],
        "Resource": [
          "arn:aws:s3:::aws-glue*"
        ]
      },
      {
        "Effect": "Allow",
        "Action": [
          "ec2:CreateTags",
          "ec2:DeleteTags"
        ],
        "Condition": {
          "ForAllValues:StringEquals": {
            "aws:TagKeys": [
              "aws-glue-service-resource"
            ]
          }
        },
        "Resource": [
                "arn:aws:ec2:*:*:network-interface/*",
                "arn:aws:ec2:*:*:security-group/*",
                "arn:aws:ec2:*:*:instance/*"
            ]
        }
    ]
}
```

The following table describes the permissions granted by this policy.

Action	Resource	Description
`"glue:*"`	`"*"`	Grants permission to run all AWS Glue API operations.
`"s3:GetBucketLocation", "s3:ListBucket", "s3:ListAllMyBuckets", "s3:GetBucketAcl"`	`"*"`	Allows listing of Amazon S3 buckets from notebook servers.
`"s3:GetObject"`	`"arn:aws:s3:::crawler-public*", "arn:aws:s3:::aws-glue-*"`	Allows get of Amazon S3 objects used by examples and tutorials from notebooks. Naming convention: Amazon S3 bucket names begin with **crawler-public** and **aws-glue-**.

Action	Resource	Description
`"s3:PutObject", "s3:DeleteObject"`	`"arn:aws:s3:::aws-glue*"`	Allows put and delete of Amazon S3 objects into your account from notebooks. Naming convention: Uses Amazon S3 folders named **aws-glue**.
`"ec2:CreateTags", "ec2:DeleteTags"`	`"arn:aws:ec2:*:*:network-interface/*", "arn:aws:ec2:*:*:security-group/*", "arn:aws:ec2:*:*:instance/*"`	Allows tagging of Amazon EC2 resources created for notebook servers. Naming convention: AWS Glue tags Amazon EC2 instances with **aws-glue-service-resource**.

5. On the **Review Policy** screen, enter your **Policy Name**, for example **GlueServiceNotebookPolicyDefault**. Enter an optional description, and when you're satisfied with the policy, choose **Create policy**.

Step 5: Create an IAM role for notebook servers

If you plan to use notebooks with development endpoints, you need to grant the IAM role permissions. You provide those permissions by using AWS Identity and Access Management IAM, through an IAM role.

> **Note**
>
> When you create an IAM role using the IAM console, the console creates an instance profile automatically and gives it the same name as the role to which it corresponds.

To create an IAM role for notebooks

1. Sign in to the AWS Management Console and open the IAM console at https://console.aws.amazon.com/iam/.
2. In the left navigation pane, choose **Roles**.
3. Choose **Create role**.
4. For role type, choose **AWS Service**, find and choose **EC2**, and choose the **EC2** use case, then choose **Next: Permissions**.
5. On the **Attach permissions policy** page, choose the policies that contain the required permissions; for example, **AWSGlueServiceNotebookRole** for general AWS Glue permissions and the AWS managed policy **AmazonS3FullAccess** for access to Amazon S3 resources. Then choose **Next: Review**.

> **Note**
>
> Ensure that one of the policies in this role grants permissions to your Amazon S3 sources and targets. Also confirm that your policy allows full access to the location where you store your notebook when you create a notebook server. You might want to provide your own policy for access to specific Amazon S3 resources. For more information about creating an Amazon S3 policy for your resources, see [Specifying Resources in a Policy]().
>
> If you plan to access Amazon S3 sources and targets that are encrypted with SSE-KMS, attach a policy that allows notebooks to decrypt the data. For more information, see [Protecting Data Using Server-Side Encryption with AWS KMS-Managed Keys (SSE-KMS)]().
>
> The following is an example.
>
> JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "kms:Decrypt"
      ],
      "Resource": [
                "arn:aws:kms:*:111122223333:key/key-id"
            ]
        }
    ]
}
```

6. For **Role name**, enter a name for your role. Create the role with the name prefixed with the string `AWSGlueServiceNotebookRole` to allow the role to be passed from console users to the notebook server. AWS Glue provided policies expect IAM service roles to begin with `AWSGlueServiceNotebookRole`. Otherwise you must add a policy to your users to allow the `iam:PassRole` permission for IAM roles to match your naming convention. For example, enter `AWSGlueServiceNotebookRoleDefault`. Then choose **Create role**.

Step 6: Create an IAM policy for SageMaker AI notebooks

If you plan to use SageMaker AI notebooks with development endpoints, you must specify permissions when you create the notebook. You provide those permissions by using AWS Identity and Access Management (IAM).

To create an IAM policy for SageMaker AI notebooks

1. Sign in to the AWS Management Console and open the IAM console at https://console.aws.amazon.com/iam/.
2. In the left navigation pane, choose **Policies**.
3. Choose **Create Policy**.
4. On the **Create Policy** page, navigate to a tab to edit the JSON. Create a policy document with the following JSON statements. Edit *`bucket-name`*, *`region-code`*, and *`account-id`* for your environment.

   JSON

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Action": [
                "s3:ListBucket"
            ],
            "Effect": "Allow",
            "Resource": [
                "arn:aws:s3:::amzn-s3-demo-bucket"
            ]
        },
        {
            "Action": [
                "s3:GetObject"
            ],
            "Effect": "Allow",
            "Resource": [
                "arn:aws:s3:::amzn-s3-demo-bucket*"
            ]
        },
        {
            "Action": [
                "logs:CreateLogStream",
                "logs:DescribeLogStreams",
                "logs:PutLogEvents",
                "logs:CreateLogGroup"
            ],
            "Effect": "Allow",
            "Resource": [
                "arn:aws:logs:us-east-1:111122223333:log-group:/aws/
sagemaker/*",
                "arn:aws:logs:us-east-1:111122223333:log-group:/aws/
sagemaker/*:log-stream:aws-glue-*"
            ]
        },
        {
            "Action": [
                "glue:UpdateDevEndpoint",
                "glue:GetDevEndpoint",
                "glue:GetDevEndpoints"
            ],
            "Effect": "Allow",
            "Resource": [
                "arn:aws:glue:us-east-1:111122223333:devEndpoint/*"
            ]
        },
        {
            "Action": [
                "sagemaker:ListTags"
            ],
                "Effect": "Allow",
                "Resource": [
                    "arn:aws:sagemaker:us-east-1:111122223333:notebook-instance/
*"
                ]
            }
        ]
}
```

Then choose **Review policy**.

The following table describes the permissions granted by this policy.

Action	Resource	Description
`"s3:ListBucket*"`	`"arn:aws:s3::: ` *`bucket-name`* ` "`	Grants permission to list Amazon S3 buckets.
`"s3:GetObject"`	`"arn:aws:s3::: ` *`bucket-name`* ` *"`	Grants permission to get Amazon S3 objects that are used by SageMaker AI notebooks.
`"logs:CreateLogStream", "logs:DescribeLogStreams", "logs:PutLogEvents", "logs:CreateLogGroup"`	`"arn:aws:logs: ` *`region-code`* ` :` *`account-id`* ` :log-group:/aws/sagemaker/*", "arn:aws:logs: ` *`region-code`* ` :` *`account-id`* ` :log-group:/aws/sagemaker/*:log-stream:aws-glue-*"`	Grants permission to write logs to Amazon CloudWatch Logs from notebooks. Naming convention: Writes to log groups whose names begin with **aws-glue**.
`"glue:UpdateDevEndpoint", "glue:GetDevEndpoint", "glue:GetDevEndpoints"`	`"arn:aws:glue: ` *`region-code`* ` :` *`account-id`* ` :devEndpoint/*"`	Grants permission to use a development endpoint from SageMaker AI notebooks.

Action	Resource	Description
`"sagemaker:ListTags"`	`"arn:aws:sagemaker: region-code :account-id :notebook-instance/*"`	Grants permission to return tags for an SageMaker AI resource. The `aws-glue-dev-endpoint` tag is required on the SageMaker AI notebook for connecting the notebook to a development endpoint.

5. On the **Review Policy** screen, enter your **Policy Name**, for example `AWSGlueSageMakerNotebook`. Enter an optional description, and when you're satisfied with the policy, choose **Create policy**.

Step 7: Create an IAM role for SageMaker AI notebooks

If you plan to use SageMaker AI notebooks with development endpoints, you need to grant the IAM role permissions. You provide those permissions by using AWS Identity and Access Management (IAM), through an IAM role.

To create an IAM role for SageMaker AI notebooks

1. Sign in to the AWS Management Console and open the IAM console at https://console.aws.amazon.com/iam/.
2. In the left navigation pane, choose **Roles**.
3. Choose **Create role**.
4. For role type, choose **AWS Service**, find and choose **SageMaker**, and then choose the **SageMaker - Execution** use case. Then choose **Next: Permissions**.
5. On the **Attach permissions policy** page, choose the policies that contain the required permissions; for example, **AmazonSageMakerFullAccess**. Choose **Next: Review**.

   If you plan to access Amazon S3 sources and targets that are encrypted with SSE-KMS, attach a policy that allows notebooks to decrypt the data, as shown in the following example. For more information, see [Protecting Data Using Server-Side Encryption with AWS KMS-Managed Keys (SSE-KMS)](#).

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "kms:Decrypt"
      ],
      "Resource": [
        "arn:aws:kms:*:111122223333:key/key-id"
      ]
    }
  ]
}
```

6. For **Role name**, enter a name for your role. To allow the role to be passed from console users to SageMaker AI, use a name that is prefixed with the string `AWSGlueServiceSageMakerNotebookRole`. AWS Glue provided policies expect IAM roles to begin with `AWSGlueServiceSageMakerNotebookRole`. Otherwise you must add a policy to your users to allow the `iam:PassRole` permission for IAM roles to match your naming convention.

   For example, enter `AWSGlueServiceSageMakerNotebookRole-Default`, and then choose **Create role**.

7. After you create the role, attach the policy that allows additional permissions required to create SageMaker AI notebooks from AWS Glue.

   Open the role that you just created, `AWSGlueServiceSageMakerNotebookRole-Default`, and choose **Attach policies**. Attach the policy that you created named `AWSGlueSageMakerNotebook` to the role.

AWS Glue access control policy examples

This section contains examples of both identity-based (IAM) access control policies and AWS Glue resource policies.

Contents

- Identity-based policy examples for AWS Glue
  - Policy best practices
  - Resource-level permissions only apply to specific AWS Glue objects
  - Using the AWS Glue console
  - Allow users to view their own permissions
  - Grant read-only permission to a table
  - Filter tables by GetTables permission
  - Grant full access to a table and all partitions
  - Control access by name prefix and explicit denial
  - Grant access using tags
  - Deny access using tags
  - Use tags with list and batch API operations
  - Control settings using condition keys or context keys
    - Control policies that control settings using condition keys
    - Control policies that control settings using context keys
  - Deny an identity the ability to create data preview sessions
- Resource-based policy examples for AWS Glue
  - Considerations for using resource-based policies with AWS Glue
  - Use a resource policy to control access in the same account

Identity-based policy examples for AWS Glue

By default, users and roles don't have permission to create or modify AWS Glue resources. They also can't perform tasks by using the AWS Management Console, AWS Command Line Interface (AWS CLI), or AWS API. To grant users permission to perform actions on the resources that they need, an IAM administrator can create IAM policies. The administrator can then add the IAM policies to roles, and users can assume the roles.

To learn how to create an IAM identity-based policy by using these example JSON policy documents, see Create IAM policies (console) in the *IAM User Guide*.

For details about actions and resource types defined by AWS Glue, including the format of the ARNs for each of the resource types, see Actions, resources, and condition keys for AWS Glue in the *Service Authorization Reference*.

> **Note**
>
> The examples provided in this section all use the `us-west-2` Region. You can replace this with the AWS Region that you want to use.

Topics

- Policy best practices
- Resource-level permissions only apply to specific AWS Glue objects
- Using the AWS Glue console
- Allow users to view their own permissions
- Grant read-only permission to a table
- Filter tables by GetTables permission
- Grant full access to a table and all partitions
- Control access by name prefix and explicit denial
- Grant access using tags
- Deny access using tags
- Use tags with list and batch API operations
- Control settings using condition keys or context keys
- Deny an identity the ability to create data preview sessions

Policy best practices

Identity-based policies determine whether someone can create, access, or delete AWS Glue resources in your account. These actions can incur costs for your AWS account. When you create or edit identity-based policies, follow these guidelines and recommendations:

- **Get started with AWS managed policies and move toward least-privilege permissions** – To get started granting permissions to your users and workloads, use the *AWS managed policies* that grant permissions for many common use cases. They are available in your AWS account. We recommend that you reduce permissions further by defining AWS customer managed policies that are specific to your use cases. For more information, see AWS managed policies or AWS managed policies for job functions in the *IAM User Guide*.

- **Apply least-privilege permissions** – When you set permissions with IAM policies, grant only the permissions required to perform a task. You do this by defining the actions that can be taken on specific resources under specific conditions, also known as *least-privilege permissions*. For more information about using IAM to apply permissions, see Policies and permissions in IAM in the *IAM User Guide*.
- **Use conditions in IAM policies to further restrict access** – You can add a condition to your policies to limit access to actions and resources. For example, you can write a policy condition to specify that all requests must be sent using SSL. You can also use conditions to grant access to service actions if they are used through a specific AWS service, such as AWS CloudFormation. For more information, see IAM JSON policy elements: Condition in the *IAM User Guide*.
- **Use IAM Access Analyzer to validate your IAM policies to ensure secure and functional permissions** – IAM Access Analyzer validates new and existing policies so that the policies adhere to the IAM policy language (JSON) and IAM best practices. IAM Access Analyzer provides more than 100 policy checks and actionable recommendations to help you author secure and functional policies. For more information, see Validate policies with IAM Access Analyzer in the *IAM User Guide*.
- **Require multi-factor authentication (MFA)** – If you have a scenario that requires IAM users or a root user in your AWS account, turn on MFA for additional security. To require MFA when API operations are called, add MFA conditions to your policies. For more information, see Secure API access with MFA in the *IAM User Guide*.

For more information about best practices in IAM, see Security best practices in IAM in the *IAM User Guide*.

Resource-level permissions only apply to specific AWS Glue objects

You can only define fine-grained control for specific objects in AWS Glue. Therefore you must write your client's IAM policy so that API operations that allow Amazon Resource Names (ARNs) for the `Resource` statement are not mixed with API operations that don't allow ARNs.

For example, the following IAM policy allows API operations for `GetClassifier` and `GetJobRun`. It defines the `Resource` as `*` because AWS Glue doesn't allow ARNs for classifiers and job runs. Because ARNs are allowed for specific API operations such as `GetDatabase` and `GetTable`, ARNs can be specified in the second half of the policy.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:GetClassifier*",
        "glue:GetJobRun*"
      ],
      "Resource": "*"
    },
    {
      "Effect": "Allow",
      "Action": [
        "glue:Get*"
      ],
      "Resource": [
        "arn:aws:glue:us-east-1:111122223333:catalog",
        "arn:aws:glue:us-east-1:111122223333:database/default",
        "arn:aws:glue:us-east-1:111122223333:table/default/e*1*",
        "arn:aws:glue:us-east-1:111122223333:connection/connection2"
      ]
    }
  ]
}
```

For a list of AWS Glue objects that allow ARNs, see Specifying AWS Glue Resource ARNs.

Using the AWS Glue console

To access the AWS Glue console, you must have a minimum set of permissions. These permissions must allow you to list and view details about the AWS Glue resources in your AWS account. If you create an identity-based policy that is more restrictive than the minimum required permissions, the console won't function as intended for entities (users or roles) with that policy.

You don't need to allow minimum console permissions for users that are making calls only to the AWS CLI or the AWS API. Instead, allow access to only the actions that match the API operation that they're trying to perform.

To ensure that users and roles can still use the AWS Glue console, also attach the AWS Glue `ConsoleAccess` or `ReadOnly` AWS managed policy to the entities. For more information, see Adding permissions to a user in the *IAM User Guide*.

For a user to work with the AWS Glue console, that user must have a minimum set of permissions that allows them to work with the AWS Glue resources for their AWS account. In addition to these AWS Glue permissions, the console requires permissions from the following services:

- Amazon CloudWatch Logs permissions to display logs.
- AWS Identity and Access Management (IAM) permissions to list and pass roles.
- AWS CloudFormation permissions to work with stacks.
- Amazon Elastic Compute Cloud (Amazon EC2) permissions to list VPCs, subnets, security groups, instances, and other objects.
- Amazon Simple Storage Service (Amazon S3) permissions to list buckets and objects, and to retrieve and save scripts.
- Amazon Redshift permissions to work with clusters.
- Amazon Relational Database Service (Amazon RDS) permissions to list instances.

For more information about the permissions that users require to view and work with the AWS Glue console, see Step 3: Attach a policy to users or groups that access AWS Glue.

If you create an IAM policy that is more restrictive than the minimum required permissions, the console won't function as intended for users with that IAM policy. To ensure that those users can still use the AWS Glue console, also attach the `AWSGlueConsoleFullAccess` managed policy as described in AWS managed (predefined) policies for AWS Glue.

Allow users to view their own permissions

This example shows how you might create a policy that allows IAM users to view the inline and managed policies that are attached to their user identity. This policy includes permissions to complete this action on the console or programmatically using the AWS CLI or AWS API.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "ViewOwnUserInfo",
            "Effect": "Allow",
            "Action": [
                "iam:GetUserPolicy",
                "iam:ListGroupsForUser",
                "iam:ListAttachedUserPolicies",
                "iam:ListUserPolicies",
                "iam:GetUser"
            ],
            "Resource": ["arn:aws:iam::*:user/${aws:username}"]
        },
        {
            "Sid": "NavigateInConsole",
            "Effect": "Allow",
            "Action": [
                "iam:GetGroupPolicy",
                "iam:GetPolicyVersion",
                "iam:GetPolicy",
                "iam:ListAttachedGroupPolicies",
                "iam:ListGroupPolicies",
                "iam:ListPolicyVersions",
                "iam:ListPolicies",
                "iam:ListUsers"
            ],
            "Resource": "*"
        }
    ]
}
```

Grant read-only permission to a table

The following policy grants read-only permission to a `books` table in database db1. For more information about resource Amazon Resource Names (ARNs), see Data Catalog ARNs.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "GetTablesActionOnBooks",
      "Effect": "Allow",
      "Action": [
        "glue:GetTables",
        "glue:GetTable"
      ],
      "Resource": [
        "arn:aws:glue:us-east-1:111122223333:catalog",
        "arn:aws:glue:us-east-1:111122223333:database/db1",
        "arn:aws:glue:us-east-1:111122223333:table/db1/books"
      ]
    }
  ]
}
```

This policy grants read-only permission to a table named `books` in the database named `db1`. To grant `Get` permission to a table, permission to the catalog and database resources is also required.

The following policy grants the minimum necessary permissions to create table `tb1` in database `db1`:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:CreateTable"
      ],
      "Resource": [
        "arn:aws:glue:us-east-1:111122223333:table/db1/tbl1",
        "arn:aws:glue:us-east-1:111122223333:database/db1",
        "arn:aws:glue:us-east-1:111122223333:catalog"
      ]
    }
  ]
}
```

Filter tables by GetTables permission

Assume that there are three tables—`customers`, `stores`, and `store_sales`—in database `db1`. The following policy grants `GetTables` permission to `stores` and `store_sales`, but not to `customers`. When you call `GetTables` with this policy, the result contains only the two authorized tables (the `customers` table is not returned).

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "GetTablesExample",
      "Effect": "Allow",
      "Action": [
        "glue:GetTables"
      ],
      "Resource": [
        "arn:aws:glue:us-east-1:111122223333:catalog",
        "arn:aws:glue:us-east-1:111122223333:database/db1",
        "arn:aws:glue:us-east-1:111122223333:table/db1/store_sales",
        "arn:aws:glue:us-east-1:111122223333:table/db1/stores"
      ]
    }
  ]
}
```

You can simplify the preceding policy by using `store*` to match any table names that start with `store`.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "GetTablesExample2",
      "Effect": "Allow",
      "Action": [
        "glue:GetTables"
      ],
      "Resource": [
        "arn:aws:glue:us-east-1:111122223333:catalog",
        "arn:aws:glue:us-east-1:111122223333:database/db1",
        "arn:aws:glue:us-east-1:111122223333:table/db1/store*"
      ]
    }
  ]
}
```

Similarly, using `/db1/*` to match all tables in `db1`, the following policy grants `GetTables` access to all the tables in `db1`.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "GetTablesReturnAll",
      "Effect": "Allow",
      "Action": [
        "glue:GetTables"
      ],
      "Resource": [
        "arn:aws:glue:us-east-1:111122223333:catalog",
        "arn:aws:glue:us-east-1:111122223333:database/db1",
        "arn:aws:glue:us-east-1:111122223333:table/db1/*"
      ]
    }
  ]
}
```

If no table ARN is provided, a call to `GetTables` succeeds, but it returns an empty list.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "GetTablesEmptyResults",
      "Effect": "Allow",
      "Action": [
        "glue:GetTables"
      ],
      "Resource": [
        "arn:aws:glue:us-east-1:111122223333:catalog",
        "arn:aws:glue:us-east-1:111122223333:database/db1"
      ]
    }
  ]
}
```

If the database ARN is missing in the policy, a call to `GetTables` fails with an `AccessDeniedException`.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "GetTablesAccessDeny",
      "Effect": "Allow",
      "Action": [
        "glue:GetTables"
      ],
      "Resource": [
        "arn:aws:glue:us-east-1:111122223333:catalog",
        "arn:aws:glue:us-east-1:111122223333:table/db1/*"
      ]
    }
  ]
}
```

Grant full access to a table and all partitions

The following policy grants all permissions on a table named `books` in database `db1`. This includes read and write permissions on the table itself, on archived versions of it, and on all its partitions.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "FullAccessOnTable",
      "Effect": "Allow",
      "Action": [
        "glue:CreateTable",
        "glue:GetTable",
        "glue:GetTables",
        "glue:UpdateTable",
        "glue:DeleteTable",
        "glue:BatchDeleteTable",
        "glue:GetTableVersion",
        "glue:GetTableVersions",
        "glue:DeleteTableVersion",
        "glue:BatchDeleteTableVersion",
        "glue:CreatePartition",
        "glue:BatchCreatePartition",
        "glue:GetPartition",
        "glue:GetPartitions",
        "glue:BatchGetPartition",
        "glue:UpdatePartition",
        "glue:DeletePartition",
        "glue:BatchDeletePartition"
      ],
      "Resource": [
        "arn:aws:glue:us-east-1:111122223333:catalog",
        "arn:aws:glue:us-east-1:111122223333:database/db1",
        "arn:aws:glue:us-east-1:111122223333:table/db1/books"
      ]
    }
  ]
}
```

The preceding policy can be simplified in practice.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "FullAccessOnTable",
      "Effect": "Allow",
      "Action": [
        "glue:*Table*",
        "glue:*Partition*"
      ],
      "Resource": [
        "arn:aws:glue:us-east-1:111122223333:catalog",
        "arn:aws:glue:us-east-1:111122223333:database/db1",
        "arn:aws:glue:us-east-1:111122223333:table/db1/books"
      ]
    }
  ]
}
```

Notice that the minimum granularity of fine-grained access control is at the table level. This means that you can't grant a user access to some partitions in a table but not others, or to some table columns but not to others. A user either has access to all of a table, or to none of it.

Control access by name prefix and explicit denial

In this example, suppose that the databases and tables in your AWS Glue Data Catalog are organized using name prefixes. The databases in the development stage have the name prefix `dev-`, and those in production have the name prefix `prod-`. You can use the following policy to grant developers full access to all databases, tables, UDFs, and so on, that have the `dev-` prefix. But you grant read-only access to everything with the `prod-` prefix.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "DevAndProdFullAccess",
      "Effect": "Allow",
      "Action": [
        "glue:*Database*",
        "glue:*Table*",
        "glue:*Partition*",
        "glue:*UserDefinedFunction*",
        "glue:*Connection*"
      ],
      "Resource": [
        "arn:aws:glue:us-east-1:111122223333:catalog",
        "arn:aws:glue:us-east-1:111122223333:database/dev-*",
        "arn:aws:glue:us-east-1:111122223333:database/prod-*",
        "arn:aws:glue:us-east-1:111122223333:table/dev-*/*",
        "arn:aws:glue:us-east-1:111122223333:table/*/dev-*",
        "arn:aws:glue:us-east-1:111122223333:table/prod-*/*",
        "arn:aws:glue:us-east-1:111122223333:table/*/prod-*",
        "arn:aws:glue:us-east-1:111122223333:userDefinedFunction/dev-*/*",
        "arn:aws:glue:us-east-1:111122223333:userDefinedFunction/*/dev-*",
        "arn:aws:glue:us-east-1:111122223333:userDefinedFunction/prod-*/*",
        "arn:aws:glue:us-east-1:111122223333:userDefinedFunction/*/prod-*",
        "arn:aws:glue:us-east-1:111122223333:connection/dev-*",
        "arn:aws:glue:us-east-1:111122223333:connection/prod-*"
      ]
    },
    {
      "Sid": "ProdWriteDeny",
      "Effect": "Deny",
      "Action": [
        "glue:*Create*",
        "glue:*Update*",
        "glue:*Delete*"
      ],
      "Resource": [
        "arn:aws:glue:us-east-1:111122223333:database/prod-*",
        "arn:aws:glue:us-east-1:111122223333:table/prod-*/*",
        "arn:aws:glue:us-east-1:111122223333:table/*/prod-*",
        "arn:aws:glue:us-east-1:111122223333:userDefinedFunction/prod-*/*",
        "arn:aws:glue:us-east-1:111122223333:userDefinedFunction/*/prod-*",
        "arn:aws:glue:us-east-1:111122223333:connection/prod-*"
      ]
    }
  ]
}
```

The second statement in the preceding policy uses explicit `deny`. You can use explicit `deny` to overwrite any `allow` permissions that are granted to the principal. This lets you lock down access to critical resources and prevent another policy from accidentally granting access to them.

In the preceding example, even though the first statement grants full access to `prod-` resources, the second statement explicitly revokes write access to them, leaving only read access to `prod-` resources.

Grant access using tags

For example, suppose that you want to limit access to a trigger `t2` to a specific user named `Tom` in your account. All other users, including `Sam`, have access to trigger `t1`. The triggers `t1` and `t2` have the following properties.

```
aws glue get-triggers
{
    "Triggers": [
        {
            "State": "CREATED",
            "Type": "SCHEDULED",
            "Name": "t1",
            "Actions": [
                {
                    "JobName": "j1"
                }
            ],
            "Schedule": "cron(0 0/1 * * ? *)"
        },
        {
            "State": "CREATED",
            "Type": "SCHEDULED",
            "Name": "t2",
            "Actions": [
                {
                    "JobName": "j1"
                }
            ],
            "Schedule": "cron(0 0/1 * * ? *)"
        }
    ]
}
```

The AWS Glue administrator attached a tag value `Tom` (`aws:ResourceTag/Name": "Tom"`) to trigger `t2`. The AWS Glue administrator also gave Tom an IAM policy with a condition statement based on the tag. As a result, Tom can only use an AWS Glue operation that acts on resources with the tag value `Tom`.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": "glue:*",
      "Resource": "*",
      "Condition": {
        "StringEquals": {
          "aws:ResourceTag/Name": "Tom"
        }
      }
    }
  ]
}
```

When Tom tries to access the trigger `t1`, he receives an access denied message. Meanwhile, he can successfully retrieve trigger `t2`.

```
aws glue get-trigger --name t1

An error occurred (AccessDeniedException) when calling the GetTrigger operation:
 User: Tom is not authorized to perform: glue:GetTrigger on resource: arn:aws:glue:us-
east-1:123456789012:trigger/t1

aws glue get-trigger --name t2
{
    "Trigger": {
        "State": "CREATED",
        "Type": "SCHEDULED",
        "Name": "t2",
        "Actions": [
            {
                "JobName": "j1"
            }
        ],
        "Schedule": "cron(0 0/1 * * ? *)"
    }
}
```

Tom can't use the plural `GetTriggers` API operation to list triggers because this operation doesn't support filtering on tags.

To give Tom access to `GetTriggers`, the AWS Glue administrator creates a policy that splits the permissions into two sections. One section allows Tom access to all triggers with the `GetTriggers` API operation. The second section allows Tom access to API operations that are tagged with the value `Tom`. With this policy, Tom is allowed both `GetTriggers` and `GetTrigger` access to trigger `t2`.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": "glue:GetTriggers",
      "Resource": "*"
    },
    {
      "Effect": "Allow",
      "Action": "glue:*",
      "Resource": "*",
      "Condition": {
        "StringEquals": {
          "aws:ResourceTag/Name": "Tom"
        }
      }
    }
  ]
}
```

Deny access using tags

Another resource policy approach is to explicitly deny access to resources.

Important

An explicit denial policy does not work for plural API operations such as `GetTriggers`.

In the following example policy, all AWS Glue job operations are allowed. However, the second `Effect` statement *explicitly* denies access to jobs tagged with the `Team` key and `Special` value.

When an administrator attaches the following policy to an identity, the identity can access all jobs *except* those tagged with the `Team` key and `Special` value.

Use tags with list and batch API operations

A third approach to writing a resource policy is to allow access to resources using a `List` API operation to list out resources for a tag value. Then, use the corresponding `Batch` API operation to allow access to details of specific resources. With this approach, the administrator doesn't need to allow access to the plural `GetCrawlers`, `GetDevEndpoints`, `GetJobs`, or `GetTriggers` API operations. Instead, you can allow the ability to list the resources with the following API operations:

- `ListCrawlers`
- `ListDevEndpoints`
- `ListJobs`
- `ListTriggers`

And, you can allow the ability to get details about individual resources with the following API operations:

- `BatchGetCrawlers`
- `BatchGetDevEndpoints`
- `BatchGetJobs`
- `BatchGetTriggers`

As an administrator, to use this approach, you can do the following:

1. Add tags to your crawlers, development endpoints, jobs, and triggers.
2. Deny user access to `Get` API operations such as `GetCrawlers`, `GetDevEndponts`, `GetJobs`, and `GetTriggers`.
3. To enable users to find out which tagged resources they have access to, allow user access to `List` API operations such as `ListCrawlers`, `ListDevEndponts`, `ListJobs`, and `ListTriggers`.
4. Deny user access to AWS Glue tagging APIs, such as `TagResource` and `UntagResource`.
5. Allow user access to resource details with `BatchGet` API operations such as `BatchGetCrawlers`, `BatchGetDevEndponts`, `BatchGetJobs`, and `BatchGetTriggers`.

For example, when calling the `ListCrawlers` operation, provide a tag value to match the user name. Then the result is a list of crawlers that match the provided tag values. Provide the list of names to `BatchGetCrawlers` to get details about each crawler with the given tag.

For example, if Tom should only be able to retrieve details of triggers that are tagged with `Tom`, the administrator can add tags to triggers for `Tom`, deny access to the `GetTriggers` API operation to all users, and allow access to all users to `ListTriggers` and `BatchGetTriggers`.

The following is the resource policy that the AWS Glue administrator grants to Tom. In the first section of the policy, AWS Glue API operations are denied for `GetTriggers`. In the second section of the policy, `ListTriggers` is allowed for all resources. However, in the third section, those resources tagged with `Tom` are allowed access with the `BatchGetTriggers` access.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Deny",
      "Action": "glue:GetTriggers",
      "Resource": "*"
    },
    {
      "Effect": "Allow",
      "Action": [
        "glue:ListTriggers"
      ],
      "Resource": [
        "*"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "glue:BatchGetTriggers"
      ],
      "Resource": [
        "*"
      ],
      "Condition": {
        "StringEquals": {
          "aws:ResourceTag/Name": "Tom"
        }
      }
    }
  ]
}
```

Using the same triggers as the previous example, Tom can access trigger `t2`, but not trigger `t1`. The following example shows the results when Tom tries to access `t1` and `t2` with `BatchGetTriggers`.

```
aws glue batch-get-triggers --trigger-names t2
{
    "Triggers": {
        "State": "CREATED",
        "Type": "SCHEDULED",
        "Name": "t2",
        "Actions": [
            {
                "JobName": "j2"
            }
        ],
        "Schedule": "cron(0 0/1 * * ? *)"
    }
}

aws glue batch-get-triggers  --trigger-names t1
```

```
An error occurred (AccessDeniedException) when calling the BatchGetTriggers operation:
 No access to any requested resource.
```

The following example shows the results when Tom tries to access both trigger `t2` and trigger `t3` (which does not exist) in the same `BatchGetTriggers` call. Notice that because Tom has access to trigger `t2` and it exists, only `t2` is returned. Although Tom is allowed to access trigger `t3`, trigger `t3` does not exist, so `t3` is returned in the response in a list of `"TriggersNotFound": []`.

```
aws glue batch-get-triggers --trigger-names t2 t3
{
    "Triggers": {
        "State": "CREATED",
        "Type": "SCHEDULED",
        "Name": "t2",
        "Actions": [
            {
                "JobName": "j2"
            }
        ],
        "TriggersNotFound": ["t3"],
        "Schedule": "cron(0 0/1 * * ? *)"
    }
}
```

Control settings using condition keys or context keys

You can use condition keys or context keys when granting permissions to create and update jobs. These sections discuss the keys:

- Control policies that control settings using condition keys
- Control policies that control settings using context keys

Control policies that control settings using condition keys

AWS Glue provides three IAM condition keys `glue:VpcIds`, `glue:SubnetIds`, and `glue:SecurityGroupIds`. You can use the condition keys in IAM policies when granting permissions to create and update jobs. You can use this setting to ensure that jobs or sessions are not created (or updated to) to run outside of a desired VPC environment. The VPC setting information is not a direct input from the `CreateJob` request, but inferred from the job "connections" field that points to an AWS Glue connection.

Example usage

Create an AWS Glue network type connection named "traffic-monitored-connection" with the desired VpcId "vpc-id1234", SubnetIds, and SecurityGroupIds.

Specify the condition keys condition for the `CreateJob` and `UpdateJob` action in the IAM policy.

```
{
  "Effect": "Allow",
  "Action": [
    "glue:CreateJob",
    "glue:UpdateJob"
  ],
  "Resource": [
    "*"
  ],
  "Condition": {
    "ForAnyValue:StringLike": {
      "glue:VpcIds": [
        "vpc-id1234"
      ]
    }
  }
}
```

You can create a similar IAM policy to prohibit creating an AWS Glue job without specifying connection information.

Restricting sessions on VPCs

To enforce created sessions to run within a specified VPC, you restrict role permission by adding a `Deny` effect on the `glue:CreateSession` action with the condition that the glue:vpc-id not equal to vpc-<123>. For example:

```
"Effect": "Deny",
"Action": [
    "glue:CreateSession"
 ],
"Condition": {
    "StringNotEquals" : {"glue:VpcIds" : ["vpc-123"]}
}
```

You also can enforce created sessions to run within a VPC by adding a `Deny` effect on the `glue:CreateSession` action with the condition that the `glue:vpc-id` is null. For example:

```
{
    "Effect": "Deny",
    "Action": [
        "glue:CreateSession"
    ],
      "Condition": {
        "Null": {"glue:VpcIds": true}
    }
},
{
    "Effect": "Allow",
    "Action": [
        "glue:CreateSession"
    ],
    "Resource": ["*"]
}
```

Control policies that control settings using context keys

AWS Glue provides a context key (`glue:CredentialIssuingService= glue.amazonaws.com`) to each role session that AWS Glue makes available to the job and developer endpoint. This allows you to implement security controls for the actions taken by AWS Glue scripts. AWS Glue provides another context key (`glue:RoleAssumedBy=glue.amazonaws.com`) to each role session where AWS Glue makes a call to another AWS service on the customer's behalf (not by a job/dev endpoint, but directly by the AWS Glue service).

Example usage

Specify the conditional permission in an IAM policy and attach it to the role to be used by an AWS Glue job. This ensures certain actions are allowed/denied based on whether the role session is used for an AWS Glue job runtime environment.

```
{
    "Effect": "Allow",
    "Action": "s3:GetObject",
    "Resource": "arn:aws:s3:::amzn-s3-demo-bucket/*",
    "Condition": {
        "StringEquals": {
            "glue:CredentialIssuingService": "glue.amazonaws.com"
        }
    }
}
```

Deny an identity the ability to create data preview sessions

This section contains an IAM policy example used to deny an identity the ability to create data preview sessions. Attach this policy to the identity, which is separate from the role used by the data preview session during its run.

```
{
    "Sid": "DatapreviewDeny",
    "Effect": "Deny",
     "Action": [
           "glue:CreateSession"
     ],
     "Resource": [
          "arn:aws:glue:*:*:session/glue-studio-datapreview*"
      ]
 }
```

Resource-based policy examples for AWS Glue

This section contains example resource-based policies, including policies that grant cross-account access.

The examples use the AWS Command Line Interface (AWS CLI) to interact with AWS Glue service API operations. You can perform the same operations on the AWS Glue console or using one of the AWS SDKs.

Important

By changing an AWS Glue resource policy, you might accidentally revoke permissions for existing AWS Glue users in your account and cause unexpected disruptions. Try these examples only in development or test accounts, and ensure that they don't break any existing workflows before you make the changes.

Topics

- Considerations for using resource-based policies with AWS Glue
- Use a resource policy to control access in the same account

Considerations for using resource-based policies with AWS Glue

> **Note**
>
> Both IAM policies and an AWS Glue resource policy take a few seconds to propagate. After you attach a new policy, you might notice that the old policy is still in effect until the new policy has propagated through the system.

You use a policy document written in JSON format to create or modify a resource policy. The policy syntax is the same as for an identity-based IAM policy (see IAM JSON policy reference), with the following exceptions:

- A `"Principal"` or `"NotPrincipal"` block is required for each policy statement.
- The `"Principal"` or `"NotPrincipal"` must identify valid existing principals. Wildcard patterns (like `arn:aws:iam::account-id:user/*`) are not allowed.
- The `"Resource"` block in the policy requires all resource ARNs to match the following regular expression syntax (where the first `%s` is the *region*, and the second `%s` is the *account-id*):

  ```
  *arn:aws:glue:%s:%s:(\*|[a-zA-Z\*]+\/?.*)
  ```

  For example, both `arn:aws:glue:us-west-2:account-id:*` and `arn:aws:glue:us-west-2:account-id:database/default` are allowed, but `*` is not allowed.
- Unlike identity-based policies, an AWS Glue resource policy must only contain Amazon Resource Names (ARNs) of resources that belong to the catalog that the policy is attached to. Such ARNs always start with `arn:aws:glue:`.
- A policy cannot cause the identity that creates it to be locked out of further policy creation or modification.
- A resource-policy JSON document cannot exceed 10 KB in size.

Use a resource policy to control access in the same account

In this example, an admin user in Account A creates a resource policy that grants IAM user `Alice` in Account A full access to the catalog. Alice has no IAM policy attached.

To do this, the admin user runs the following AWS CLI command.

```
# Run as admin of Account A
$ aws glue put-resource-policy --profile administrator-name --region us-west-2 --
policy-in-json '{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Principal": {
        "AWS": [
          "arn:aws:iam::account-A-id:user/Alice"
        ]
      },
      "Effect": "Allow",
      "Action": [
        "glue:*"
      ],
      "Resource": [
        "arn:aws:glue:us-west-2:account-A-id:*"
      ]
    }
  ]
}'
```

Instead of entering the JSON policy document as a part of your AWS CLI command, you can save a policy document in a file and reference the file path in the AWS CLI command, prefixed by `file://`. The following is an example of how you might do that.

```
$ echo '{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Principal": {
        "AWS": [
          "arn:aws:iam::account-A-id:user/Alice"
        ]
      },
      "Effect": "Allow",
      "Action": [
        "glue:*"
      ],
      "Resource": [
        "arn:aws:glue:us-west-2:account-A-id:*"
      ]
    }
  ]
}' > /temp/policy.json

$ aws glue put-resource-policy --profile admin1 \
    --region us-west-2 --policy-in-json file:///temp/policy.json
```

After this resource policy has propagated, Alice can access all AWS Glue resources in Account A, as follows.

```
# Run as user Alice
$ aws glue create-database --profile alice --region us-west-2 --database-input '{
    "Name": "new_database",
    "Description": "A new database created by Alice",
    "LocationUri": "s3://amzn-s3-demo-bucket"
}'

$ aws glue get-table --profile alice --region us-west-2 --database-name "default" --
table-name "tbl1"}
```

In response to Alice's `get-table` call, the AWS Glue service returns the following.

```
{
  "Table": {
    "Name": "tbl1",
    "PartitionKeys": [],
    "StorageDescriptor": {
        ......
    },
    ......
  }
}
```

Granting AWS managed policies for AWS Glue

An AWS managed policy is a standalone policy that is created and administered by AWS. AWS managed policies are designed to provide permissions for many common use cases so that you can start assigning permissions to users, groups, and roles.

Keep in mind that AWS managed policies might not grant least-privilege permissions for your specific use cases because they're available for all AWS customers to use. We recommend that you reduce permissions further by defining customer managed policies that are specific to your use cases.

You cannot change the permissions defined in AWS managed policies. If AWS updates the permissions defined in an AWS managed policy, the update affects all principal identities (users, groups, and roles) that the policy is attached to. AWS is most likely to update an AWS managed policy when a new AWS service is launched or new API operations become available for existing services.

For more information, see AWS managed policies in the *IAM User Guide*.

AWS managed (predefined) policies for AWS Glue

AWS addresses many common use cases by providing standalone IAM policies that are created and administered by AWS. These AWS managed policies grant necessary permissions for common use cases so that you can avoid having to investigate what permissions are needed. For more information, see AWS managed policies in the *IAM User Guide*.

The following AWS managed policies, which you can attach to identities in your account, are specific to AWS Glue and are grouped by use case scenario:

- AWSGlueConsoleFullAccess – Grants full access to AWS Glue resources when an identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console.
- AWSGlueServiceRole – Grants access to resources that various AWS Glue processes require to run on your behalf. These resources include AWS Glue, Amazon S3, IAM, CloudWatch Logs, and Amazon EC2. If you follow the naming convention for resources specified in this policy, AWS Glue processes have the required permissions. This policy is typically attached to roles specified when defining crawlers, jobs, and development endpoints.

- AwsGlueSessionUserRestrictedServiceRole – Provides full access to all AWS Glue resources except for sessions. It allows users to create and use only the interactive sessions that are associated with the user. This policy includes other permissions needed by AWS Glue to manage AWS Glue resources in other AWS services. The policy also allows adding tags to AWS Glue resources in other AWS services.

  > **Note**
  >
  > To achieve the full security benefits, do not grant this policy to a user that was assigned the `AWSGlueServiceRole`, `AWSGlueConsoleFullAccess`, or `AWSGlueConsoleSageMakerNotebookFullAccess` policy.

- AwsGlueSessionUserRestrictedPolicy – Provides access to create AWS Glue interactive sessions using the `CreateSession` API operation only if a tag key “owner” and value that match the assignee's AWS user ID are provided. This identity policy is attached to the IAM user that invokes the `CreateSession` API operation. This policy also permits the assignee to interact with the AWS Glue interactive session resources that were created with an “owner” tag and value that match their AWS user ID. This policy denies permission to change or remove "owner" tags from an AWS Glue session resource after the session is created.

  > **Note**
  >
  > To achieve the full security benefits, do not grant this policy to a user that was assigned the `AWSGlueServiceRole`, `AWSGlueConsoleFullAccess`, or `AWSGlueConsoleSageMakerNotebookFullAccess` policy.

- AwsGlueSessionUserRestrictedNotebookServiceRole – Provides sufficient access to the AWS Glue Studio notebook session to interact with specific AWS Glue interactive session resources. These are resources that are created with the “owner” tag value that matches the AWS user ID of the principal (IAM user or role) that creates the notebook. For more information about these tags, see the Principal key values chart in the *IAM User Guide*.

  This service-role policy is attached to the role that is specified with a magic command within the notebook or is passed as a role to the `CreateSession` API operation. This policy also permits the principal to create an AWS Glue interactive session from the AWS Glue Studio notebook interface only if a tag key “owner” and value match the AWS user ID of the principal. This policy denies permission to change or remove "owner" tags from an AWS Glue session resource after the session is created. This policy also includes permissions for writing and reading from Amazon

S3 buckets, writing CloudWatch logs, and creating and deleting tags for Amazon EC2 resources used by AWS Glue.

> **Note**
>
> To achieve the full security benefits, do not grant this policy to a role that was assigned the `AWSGlueServiceRole`, `AWSGlueConsoleFullAccess`, or `AWSGlueConsoleSageMakerNotebookFullAccess` policy.

- AwsGlueSessionUserRestrictedNotebookPolicy – Provides access to create an AWS Glue interactive session from the AWS Glue Studio notebook interface only if there is a tag key “owner” and value that match the AWS user IDof the principal (IAM user or role) that creates the notebook. For more information about these tags, see the Principal key values chart in the *IAM User Guide*.

  This policy is attached to the principal (IAM user or role) that creates sessions from the AWS Glue Studio notebook interface. This policy also permits sufficient access to the AWS Glue Studio notebook to interact with specific AWS Glue interactive session resources. These are resources that are created with the “owner” tag value that matches the AWS user ID of the principal. This policy denies permission to change or remove "owner" tags from an AWS Glue session resource after the session is created.
- AWSGlueServiceNotebookRole – Grants access to AWS Glue sessions started in an AWS Glue Studio notebook. This policy allows listing and getting session information for all sessions, but only permits users to create and use the sessions tagged with their AWS user ID. This policy denies permission to change or remove “owner” tags from AWS Glue session resources tagged with their AWS ID.

  Assign this policy to the AWS user who creates jobs using the notebook interface in AWS Glue Studio.
- AWSGlueConsoleSageMakerNotebookFullAccess – Grants full access to AWS Glue and SageMaker AI resources when the identity that the policy is attached to uses the AWS Management Console. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console who manage SageMaker AI notebooks.
- AWSGlueSchemaRegistryFullAccess – Grants full access to AWS Glue Schema Registry resources when the identity that the policy is attached to uses the AWS Management Console or AWS CLI. If you follow the naming convention for resources specified in this policy, users have full console

capabilities. This policy is typically attached to users of the AWS Glue console or AWS CLI who manage the AWS Glue Schema Registry.

- AWSGlueSchemaRegistryReadonlyAccess – Grants read-only access to AWS Glue Schema Registry resources when an identity that the policy is attached to uses the AWS Management Console or AWS CLI. If you follow the naming convention for resources specified in this policy, users have full console capabilities. This policy is typically attached to users of the AWS Glue console or AWS CLI who use the AWS Glue Schema Registry.

> **Note**
>
> You can review these permissions policies by signing in to the IAM console and searching for specific policies there.

You can also create your own custom IAM policies to allow permissions for AWS Glue actions and resources. You can attach these custom policies to the IAM users or groups that require those permissions.

To create a connection with VPC configuration while using a custom IAM role, it must have the following VPC access actions:

- secretsmanager:GetSecretValue
- secretsmanager:PutSecretValue
- secretsmanager:DescribeSecret
- ec2:CreateNetworkInterface
- ec2:DeleteNetworkInterface
- ec2:DescribeNetworkInterfaces
- ec2:DescribeSubnets

AWS Glue updates to AWS managed policies

View details about updates to AWS managed policies for AWS Glue since this service began tracking these changes. For automatic alerts about changes to this page, subscribe to the RSS feed on the AWS Glue Document history page.

Change	Description	Date
AwsGlueSessionUserRestrictedNotebookPolicy – Minor update to an existing policy.	Add allow for `glue:TagResource` action on the owner tag key. Required for supporting on tag-on-create for sessions with owner tag key.	August 30, 2024
AwsGlueSessionUserRestrictedNotebookServiceRole – Minor update to an existing policy.	Add allow for `glue:TagResource` action on the owner tag key. Required for supporting on tag-on-create for sessions with owner tag key.	August 30, 2024
AwsGlueSessionUserRestrictedPolicy – Minor update to an existing policy.	Add allow for `glue:TagResource` action on the owner tag key. Required for supporting on tag-on-create for sessions with owner tag key.	August 5, 2024
AwsGlueSessionUserRestrictedServiceRole – Minor update to an existing policy.	Add allow for `glue:TagResource` action on the owner tag key. Required for supporting on tag-on-create for sessions with owner tag key.	August 5, 2024
AwsGlueSessionUserRestrictedPolicy – Minor update to an existing policy.	Add `glue:StartCompletion` and `glue:GetCompletion` to policy. Required for Amazon Q data integration in AWS Glue.	April 30, 2024
AwsGlueSessionUserRestrictedNotebookServiceRole –	Add `glue:StartCompletion` and `glue:GetC`	April 30, 2024

Change	Description	Date
Minor update to an existing policy.	`ompletion` to policy. Required for Amazon Q data integration in AWS Glue.	
AwsGlueSessionUserRestrictedServiceRole – Minor update to an existing policy.	Add `glue:StartCompletion` and `glue:GetCompletion` to policy. Required for Amazon Q data integration in AWS Glue.	April 30, 2024
AWSGlueServiceNotebookRole – Minor update to an existing policy.	Add `glue:StartCompletion` and `glue:GetCompletion` to policy. Required for Amazon Q data integration in AWS Glue.	Jan 30, 2024
AwsGlueSessionUserRestrictedNotebookPolicy – Minor update to an existing policy.	Add `glue:StartCompletion` and `glue:GetCompletion` to policy. Required for Amazon Q data integration in AWS Glue.	Nov 29, 2023
AWSGlueServiceNotebookRole – Minor update to an existing policy.	Add `codewhisperer:GenerateRecommendations` to policy. Required for a new feature where AWS Glue generates CodeWhisperer recommendations.	Oct 9, 2023
AWSGlueServiceRole – Minor update to an existing policy.	Tighten scope of CloudWatch permissions to better reflect AWS Glue logging.	Aug 4, 2023

Change	Description	Date
AWSGlueConsoleFullAccess – Minor update to an existing policy.	Add `databrew` recipe List and Describe permissions to policy. Required to provide full administrative access for new features where AWS Glue can access recipes.	May 9, 2023
AWSGlueConsoleFullAccess – Minor update to an existing policy.	Add `cloudformation:ListStacks` to policy. Preserves existing capabilities after changes to AWS CloudFormation authorization requirements.	March 28, 2023
New managed policies added for the interactive sessions feature: - AwsGlueSessionUserRestrictedServiceRole - AwsGlueSessionUserRestrictedPolicy - AwsGlueSessionUserRestrictedNotebookServiceRole - AwsGlueSessionUserRestrictedNotebookPolicy	These policies were designed to provide additional security for interactive sessions and notebooks in AWS Glue Studio. The policies restrict access to the `CreateSession` API operation so that only the owner has access.	November 30, 2021

Change	Description	Date
AWSGlueConsoleSageMakerNotebookFullAccess – Update to an existing policy.	Removed a redundant resource ARN (`arn:aws:s3:::aws-glue-*/*`) for the action that grants read/write permissions on Amazon S3 buckets that AWS Glue uses to store scripts and temporary files. Fixed a syntax issue by changing `"StringEquals"` to `"ForAnyValue:StringLike"`, and moved the `"Effect": "Allow"` lines to precede the `"Action":` line in each place where they were out of order.	July 15, 2021
AWSGlueConsoleFullAccess – Update to an existing policy.	Removed a redundant resource ARN (`arn:aws:s3:::aws-glue-*/*`) for the action that grants read/write permissions on Amazon S3 buckets that AWS Glue uses to store scripts and temporary files.	July 15, 2021
AWS Glue started tracking changes.	AWS Glue started tracking changes for its AWS managed policies.	June 10, 2021

Granting dynamically scoped policies for job execution

AWS Glue offers a powerful new capability: dynamic session policies for job executions. This feature allows you to specify custom, fine-grained permissions for each job run without creating multiple IAM roles.

When you start a Glue job using the `StartJobRun` API, you can include an inline session policy. This policy temporarily modifies the permissions of the job's execution role for the duration of that specific job run. It's similar to using temporary credentials with the `AssumeRole` API in other AWS services.

- **Enhanced security**: You can limit job permissions to the minimum necessary for each run.
- **Simplified management**: Eliminates the need to create and maintain numerous IAM roles for different scenarios.
- **Flexibility**: You can be adjust permissions dynamically based on runtime parameters or tenant-specific needs.
- **Scalability**: This method excels in multi-tenant environments where you need to isolate resources between tenants.

Examples for granting dynamically scoped policies usage:

The following examples demonstrate granting jobs *read* and *write* access only to a specific Amazon S3 bucket path, where the path is dynamically determined by the job run ID. This illustrates how to implement granular, execution-specific permissions for each job run.

From CLI

```
aws glue start-job-run \
    --job-name "your-job-name" \
    --execution-role-session-policy '{
        "Version": "2012-10-17",
        "Statement": [
            {
                "Effect": "Allow",
                "Action": [
                    "s3:GetObject",
                    "s3:PutObject"
                ],
                "Resource": [
                        "arn:aws:s3:::specific-bucket/${JobRunId}/*"
                    ]
                }
            ]
        }'
```

Specifying AWS Glue resource ARNs

In AWS Glue, you can control access to resources using an AWS Identity and Access Management (IAM) policy. In a policy, you use an Amazon Resource Name (ARN) to identify the resource that the policy applies to. Not all resources in AWS Glue support ARNs.

Topics

- Data Catalog ARNs
- ARNs for non-catalog objects in AWS Glue
- Access control for AWS Glue non-catalog singular API operations
- Access control for AWS Glue non-catalog API operations that retrieve multiple items
- Access control for AWS Glue non-catalog BatchGet API operations

Data Catalog ARNs

Data Catalog resources have a hierarchical structure, with `catalog` as the root.

```
arn:aws:glue:region:account-id:catalog
```

Each AWS account has a single Data Catalog in an AWS Region with the 12-digit account ID as the catalog ID. Resources have unique ARNs associated with them, as shown in the following table.

Resource type	ARN format
Catalog	`arn:aws:glue: region:account-id :catalog` For example: `arn:aws:glue:us-east-1:123456789012:catalog`
Database	`arn:aws:glue: region:account-id :database/ database name`

Resource type	ARN format
	For example: `arn:aws:glue:us-east-1:123456789012:database/db1`
Table	`arn:aws:glue: *region*:*account-id* :table/*database name*/*table name*` For example: `arn:aws:glue:us-east-1:123456789012:table/db1/tbl1`
Federated S3 tables catalog (All table buckets)	`arn:aws:glue: *region*:*account-id* :catalog/s3tablescatalog` For example: `arn:aws:glue:us-east-1:123456789012:catalog/s3tablescatalog`
Federated S3 table bucket catalog (child catalog)	`arn:aws:glue: *region*:*account-id* :catalog/s3tablescatalog/ *bucket name*` For example: `arn:aws:glue:us-east-1:123456789012:catalog/s3tablescatalog/amzn-s3-demo-bucket1`
Federated S3 tables database	`arn:aws:glue: *region*:*account-id* :database/s3tablescatalog/ *child catalog name*/*database name*` For example: `arn:aws:glue:us-east-1:123456789012:database/s3tablescatalog/amzn-s3-demo-bucket1/nsdb1`
Federated S3 table	`arn:aws:glue: *region*:*account-id* :table/s3tablescatalog/ *child catalog name*/*database name*/*table name*` For example: `arn:aws:glue:us-east-1:123456789012:table/s3tablescatalog/amzn-s3-demo-bucket1/nsdb1/s3tbl1`

Resource type	ARN format
Federated S3 tables catalog (A single table bucket registered with Lake Formation)	`arn:aws:glue:` *`region`*`:`*`account-id`* `:catalog/`*`catalog name`* For example: `arn:aws:glue:us-east-1:123456789012: catalog/amzn-s3-demo-bucket1`
Federated S3 tables database	`arn:aws:glue:` *`region`*`:`*`account-id`* `:catalog/`*`catalog name`*`/`*`database name`* For example: `arn:aws:glue:us-east-1:123456789012: database/amzn-s3-demo-bucket1/nsdb1`
Federated S3 table	`arn:aws:glue:` *`region`*`:`*`account-id`* `:catalog/`*`catalog name`*`/`*`database name`*`/`*`table name`* For example: `arn:aws:glue:us-east-1:123456789012: table/amzn-s3-demo-bucket1/` *`nsdb1`*`/s3tbl1`
Federated or managed catalog (top-level catalog in a multi-catalog)	`arn:aws:glue:` *`region`*`:`*`account-id`* `:catalog/` *`top-level catalog name`* For example: `arn:aws:glue:us-east-1:123456789012: catalog/nscatalog` ARN format for a managed catalog follows the same structure.
Federated multi-level catalog (child catalog in a multi-level catalog)	`arn:aws:glue:` *`region`*`:`*`account-id`* `:catalog/` *`top-level catalog name`*`/`*`child catalog name`* For example: `arn:aws:glue:us-east-1:123456789012: catalog/nscatalog/dbcatalog`
Federated database	`arn:aws:glue:` *`region`*`:`*`account-id`* `:database/` *`name space catalog name`*`/`*`child catalog name`*`/`*`database name`* For example: `arn:aws:glue:us-east-1:123456789012: database/nscatalog/dbcatalog/schemadb`

Resource type	ARN format
Federated table	`arn:aws:glue:` *`region`*`:`*`account-id`* `:table/`*`name space catalog name`*`/`*`child catalog name`*`/`*`database name`*`/`*`table name`* For example: `arn:aws:glue:us-east-1:123456789012:table/nscatalog/dbcatalog/schemadb/rstbl1`
Catalog link container	`arn:aws:glue:` *`region`*`:`*`account-id`* `:catalog/`*`link container name`* For example: `arn:aws:glue:glue:us-east-1:123456789012:catalog` /linkcontainer-example
Database	`arn:aws:glue:` *`region`*`:`*`account-id`* `:catalog/`*`link container name`*`/`*`database name`* For example: `arn:aws:glue:glue:us-east-1:123456789012:database` /linkcontainer-example/link-db
User-defined function	`arn:aws:glue:` *`region`*`:`*`account-id`* `:userDefinedFunction/` *`database name`*`/`*`user-defined function name`* For example: `arn:aws:glue:us-east-1:123456789012:userDefinedFunction/db1/func1`
Connection	`arn:aws:glue:` *`region`*`:`*`account-id`* `:connection/` *`connection name`* For example: `arn:aws:glue:us-east-1:123456789012:connection/connection1`
Interactive Session	`arn:aws:glue:` *`region`*`:`*`account-id`* `:session/` *`interactive session id`* For example: `arn:aws:glue:us-east-1:123456789012:session/a1b2c3d4-5678-90ab-cdef-EXAMPLE11111`

To enable fine-grained access control, you can use these ARNs in your IAM policies and resource policies to grant and deny access to specific resources. Wildcards are allowed in the policies. For example, the following ARN matches all tables in database `default`.

```
arn:aws:glue:us-east-1:123456789012:table/default/*
```

Important

All operations performed on a Data Catalog resource require permission on the resource and all the ancestors of that resource. For example, to create a partition for a table requires permission on the table, database, and catalog where the table is located. The following example shows the permission required to create partitions on table `PrivateTable` in database `PrivateDatabase` in the Data Catalog.

```
{
   "Sid": "GrantCreatePartitions",
   "Effect": "Allow",
   "Action": [
       "glue:BatchCreatePartitions"
   ],
   "Resource": [
       "arn:aws:glue:us-east-1:123456789012:table/PrivateDatabase/PrivateTable",
       "arn:aws:glue:us-east-1:123456789012:database/PrivateDatabase",
       "arn:aws:glue:us-east-1:123456789012:catalog"
   ]
}
```

In addition to permission on the resource and all its ancestors, all delete operations require permission on all children of that resource. For example, deleting a database requires permission on all the tables and user-defined functions in the database, in addition to the database and the catalog where the database is located. The following example shows the permission required to delete database `PrivateDatabase` in the Data Catalog.

```
{
   "Sid": "GrantDeleteDatabase",
   "Effect": "Allow",
   "Action": [
       "glue:DeleteDatabase"
   ],
   "Resource": [
        "arn:aws:glue:us-east-1:123456789012:table/PrivateDatabase/*",
        "arn:aws:glue:us-east-1:123456789012:userDefinedFunction/PrivateDatabase/
*",
        "arn:aws:glue:us-east-1:123456789012:database/PrivateDatabase",
        "arn:aws:glue:us-east-1:123456789012:catalog"
    ]
}
```

In summary, actions on Data Catalog resources follow these permission rules:

- Actions on the catalog require permission on the catalog only.
- Actions on a database require permission on the database and catalog.
- Delete actions on a database require permission on the database and catalog plus all tables and user-defined functions in the database.
- Actions on a table, partition, or table version require permission on the table, database, and catalog.
- Actions on a user-defined function require permission on the user-defined function, database, and catalog.
- Actions on a connection require permission on the connection and catalog.

ARNs for non-catalog objects in AWS Glue

Some AWS Glue resources allow resource-level permissions to control access using an ARN. You can use these ARNs in your IAM policies to enable fine-grained access control. The following table lists the resources that can contain resource ARNs.

Resource type	ARN format
Crawler	`arn:aws:glue:` *`region`*`:`*`account-id`* `:crawler/` *`crawler-name`* For example: `arn:aws:glue:us-east-1:123456789012:crawler/mycrawler`
Job	`arn:aws:glue:` *`region`*`:`*`account-id`* `:job/`*`job-name`*

Resource type	ARN format
	For example: `arn:aws:glue:us-east-1:123456789012:job/testjob`
Trigger	`arn:aws:glue:` *`region`*`:`*`account-id`* `:trigger/` *`trigger-name`* For example: `arn:aws:glue:us-east-1:123456789012:trigger/sampletrigger`
Development endpoint	`arn:aws:glue:` *`region`*`:`*`account-id`* `:devEndpoint/` *`development-endpoint-name`* For example: `arn:aws:glue:us-east-1:123456789012:devEndpoint/temporarydevendpoint`
Machine learning transform	`arn:aws:glue:` *`region`*`:`*`account-id`* `:mlTransform/` *`transform-id`* For example: `arn:aws:glue:us-east-1:123456789012:mlTransform/tfm-1234567890`

Access control for AWS Glue non-catalog singular API operations

AWS Glue non-catalog *singular* API operations act on a single item (development endpoint). Examples are `GetDevEndpoint`, `CreateUpdateDevEndpoint`, and `UpdateDevEndpoint`. For these operations, a policy must put the API name in the `"action"` block and the resource ARN in the `"resource"` block.

Suppose that you want to allow a user to call the `GetDevEndpoint` operation. The following policy grants the minimum necessary permissions to an endpoint named `myDevEndpoint-1`.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "MinimumPermissions",
      "Effect": "Allow",
      "Action": "glue:GetDevEndpoint",
      "Resource": "arn:aws:glue:us-east-1:111122223333:devEndpoint/
myDevEndpoint-1"
    }
  ]
}
```

The following policy allows `UpdateDevEndpoint` access to resources that match `myDevEndpoint-` with a wildcard (*).

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "PermissionWithWildcard",
      "Effect": "Allow",
      "Action": "glue:UpdateDevEndpoint",
      "Resource": "arn:aws:glue:us-east-1:111122223333:devEndpoint/myDevEndpoint-
*"
    }
  ]
}
```

You can combine the two policies as in the following example. You might see `EntityNotFoundException` for any development endpoint whose name begins with `A`. However, an access denied error is returned when you try to access other development endpoints.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "CombinedPermissions",
      "Effect": "Allow",
            "Action": [
              "glue:UpdateDevEndpoint",
              "glue:GetDevEndpoint"
            ],
            "Resource": "arn:aws:glue:us-east-1:111122223333:devEndpoint/A*"
          }
        ]
      }
```

Access control for AWS Glue non-catalog API operations that retrieve multiple items

Some AWS Glue API operations retrieve multiple items (such as multiple development endpoints); for example, `GetDevEndpoints`. For this operation, you can specify only a wildcard (*) resource, and not specific ARNs.

For example, to include `GetDevEndpoints` in the policy, the resource must be scoped to the wildcard (*). The singular operations (`GetDevEndpoint`, `CreateDevEndpoint`, and `DeleteDevendpoint`) are also scoped to all (*) resources in the example.

```
{
            "Sid": "PluralAPIIncluded",
            "Effect": "Allow",
            "Action": [
                "glue:GetDevEndpoints",
                "glue:GetDevEndpoint",
                "glue:CreateDevEndpoint",
                "glue:UpdateDevEndpoint"
            ],
            "Resource": [
                "*"
            ]
}
```

Access control for AWS Glue non-catalog BatchGet API operations

Some AWS Glue API operations retrieve multiple items (such as multiple development endpoints); for example, `BatchGetDevEndpoints`. For this operation, you can specify an ARN to limit the scope of resources that can be accessed.

For example, to allow access to a specific development endpoint, include `BatchGetDevEndpoints` in the policy with its resource ARN.

```
{
            "Sid": "BatchGetAPIIncluded",
            "Effect": "Allow",
            "Action": [
                "glue:BatchGetDevEndpoints"
            ],
            "Resource": [
                "arn:aws:glue:us-east-1:123456789012:devEndpoint/de1"
            ]
}
```

With this policy, you can successfully access the development endpoint named `de1`. However, if you try to access the development endpoint named `de2`, an error is returned.

```

An error occurred (AccessDeniedException) when calling the BatchGetDevEndpoints
 operation: No access to any requested resource.
```

⚠ **Important**

For alternative approaches to setting up IAM policies, such as using `List` and `BatchGet` API operations, see Identity-based policy examples for AWS Glue.

Granting cross-account access

Granting access to Data Catalog resources across accounts enables your extract, transform, and load (ETL) jobs to query and join data from different accounts.

Topics

- Methods for granting cross-account access in AWS Glue
- Adding or updating the Data Catalog resource policy
- Making a cross-account API call
- Making a cross-account ETL call
- Cross-account CloudTrail logging

- Cross-account resource ownership and billing
- Cross-account access limitations

Methods for granting cross-account access in AWS Glue

You can grant access to your data to external AWS accounts by using AWS Glue methods or by using AWS Lake Formation cross-account grants. The AWS Glue methods use AWS Identity and Access Management (IAM) policies to achieve fine-grained access control. Lake Formation uses a simpler `GRANT/REVOKE` permissions model similar to the `GRANT/REVOKE` commands in a relational database system.

This section describes using the AWS Glue methods. For information about using Lake Formation cross-account grants, see Granting Lake Formation Permissions in the *AWS Lake Formation Developer Guide*.

There are two AWS Glue methods for granting cross-account access to a resource:

- Use a Data Catalog resource policy
- Use an IAM role

Granting cross-account access using a resource policy

The following are the general steps for granting cross-account access using a Data Catalog resource policy:

1. An administrator (or other authorized identity) in Account A attaches a resource policy to the Data Catalog in Account A. This policy grants Account B specific cross-account permissions to perform operations on a resource in Account A's catalog.
2. An administrator in Account B attaches an IAM policy to an IAM identity in Account B that delegates the permissions received from Account A.

   The identity in Account B now has access to the specified resource in Account A.

   The identity needs permission from *both* the resource owner (Account A) *and* their parent account (Account B) to be able to access the resource.

Granting cross-account access using an IAM role

The following are the general steps for granting cross-account access using an IAM role:

1. An administrator (or other authorized identity) in the account that owns the resource (Account A) creates an IAM role.
2. The administrator in Account A attaches a policy to the role that grants cross-account permissions for access to the resource in question.
3. The administrator in Account A attaches a trust policy to the role that identifies an IAM identity in a different account (Account B) as the principal who can assume the role.

   The principal in the trust policy can also be an AWS service principal if you want to grant an AWS service permission to assume the role.
4. An administrator in Account B now delegates permissions to one or more IAM identities in Account B so that they can assume that role. Doing so gives those identities in Account B access to the resource in account A.

For more information about using IAM to delegate permissions, see Access management in the *IAM User Guide*. For more information about users, groups, roles, and permissions, see Identities (users, groups, and roles) in the *IAM User Guide*.

For a comparison of these two approaches, see How IAM roles differ from resource-based policies in the *IAM User Guide*. AWS Glue supports both options, with the restriction that a resource policy can grant access only to Data Catalog resources.

For example, to give the `Dev` role in Account B access to database `db1` in Account A, attach the following resource policy to the catalog in Account A.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:GetDatabase"
      ],
      "Principal": {
        "AWS": [
          "arn:aws:iam::111122223333:role/Dev"
        ]
      },
      "Resource": [
        "arn:aws:glue:us-east-1:111122223333:catalog",
        "arn:aws:glue:us-east-1:111122223333:database/db1"
      ]
    }
  ]
}
```

In addition, Account B would have to attach the following IAM policy to the `Dev` role before it would actually get access to db1 in Account A.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:GetDatabase"
      ],
      "Resource": [
        "arn:aws:glue:us-east-1:111122223333:catalog",
        "arn:aws:glue:us-east-1:111122223333:database/db1"
      ]
    }
  ]
}
```

Adding or updating the Data Catalog resource policy

You can add or update the AWS Glue Data Catalog resource policy using the console, API, or AWS Command Line Interface (AWS CLI).

Important

If you have already made cross-account permission grants from your account with AWS Lake Formation, adding or updating the Data Catalog resource policy requires an extra step. For more information, see Managing cross-account permissions using both AWS Glue and Lake Formation in the *AWS Lake Formation Developer Guide*.

To determine if Lake Formation cross-account grants exist, use the `glue:GetResourcePolicies` API operation or the AWS CLI. If `glue:GetResourcePolicies` returns any policies other than an already existing Data Catalog policy, then Lake Formation grants exist. For more information, see Viewing all cross-account grants using the GetResourcePolicies API operation in the *AWS Lake Formation Developer Guide*.

To add or update the Data Catalog resource policy (console)

1. Open the AWS Glue console at https://console.aws.amazon.com/glue/.

   Sign in as an AWS Identity and Access Management (IAM) administrative user who has the `glue:PutResourcePolicy` permission.
2. In the navigation pane, choose **Settings**.
3. On the **Data catalog settings** page, under **Permissions**, paste a resource policy into the text area. Then choose **Save**.

   If the console displays a alert stating that the permissions in the policy will be in addition to any permissions granted using Lake Formation, choose **Proceed**.

To add or update the Data Catalog resource policy (AWS CLI)

- Submit an `aws glue put-resource-policy` command. If Lake Formation grants already exist, ensure that you include the `--enable-hybrid` option with the value `'TRUE'`.

  For examples of using this command, see Resource-based policy examples for AWS Glue.

Making a cross-account API call

All AWS Glue Data Catalog operations have a `CatalogId` field. If the required permissions have been granted to enable cross-account access, a caller can make Data Catalog API calls across accounts. The caller does this by passing the target AWS account ID in `CatalogId` so as to access the resource in that target account.

If no `CatalogId` value is provided, AWS Glue uses the caller's own account ID by default, and the call is not cross-account.

Making a cross-account ETL call

Some AWS Glue PySpark and Scala APIs have a catalog ID field. If all the required permissions have been granted to enable cross-account access, an ETL job can make PySpark and Scala calls to API operations across accounts by passing the target AWS account ID in the catalog ID field to access Data Catalog resources in a target account.

If no catalog ID value is provided, AWS Glue uses the caller's own account ID by default, and the call is not cross-account.

For PySpark APIs that support `catalog_id`, see GlueContext class. For Scala APIs that support `catalogId`, see AWS Glue Scala GlueContext APIs.

The following example shows the permissions required by the grantee to run an ETL job. In this example, *`grantee-account-id`* is the `catalog-id` of the client running the job and *`grantor-account-id`* is the owner of the resource. This example grants permission to all catalog resources in the grantor's account. To limit the scope of resources granted, you can provide specific ARNs for the catalog, database, table, and connection.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:GetConnection",
        "glue:GetDatabase",
          "glue:GetTable",
          "glue:GetPartition"
        ],
        "Principal": {
          "AWS": [
            "arn:aws:iam::111122223333:root"
          ]
        },
        "Resource": [
          "arn:aws:glue:us-east-1:111122223333:*"
        ]
      }
    ]
  }
```

> **Note**
>
> If a table in the grantor's account points to an Amazon S3 location that is also in the grantor's account, the IAM role used to run an ETL job in the grantee's account must have permission to list and get objects from the grantor's account.

Given that the client in Account A already has permission to create and run ETL jobs, the following are the basic steps to set up an ETL job for cross-account access:

1. Allow cross-account data access (skip this step if Amazon S3 cross-account access is already set up).
    a. Update the Amazon S3 bucket policy in Account B to allow cross-account access from Account A.
    b. Update the IAM policy in Account A to allow access to the bucket in Account B.
2. Allow cross-account Data Catalog access.
    a. Create or update the resource policy attached to the Data Catalog in Account B to allow access from Account A.
    b. Update the IAM policy in Account A to allow access to the Data Catalog in Account B.

Cross-account CloudTrail logging

When an AWS Glue extract, transform, and load (ETL) job accesses the underlying data of a Data Catalog table shared through AWS Lake Formation cross-account grants, there is additional AWS CloudTrail logging behavior.

For purposes of this discussion, the AWS account that shared the table is the owner account, and the account that the table was shared with is the recipient account. When an ETL job in the recipient account accesses data in the table in the owner account, the data-access CloudTrail event that is added to the logs for the recipient account gets copied to the owner account's CloudTrail logs. This is so owner accounts can track data accesses by the various recipient accounts. By default, the CloudTrail events do not include a human-readable principal identifier (principal ARN). An administrator in the recipient account can opt in to include the principal ARN in the logs.

For more information, see Cross-account CloudTrail logging in the *AWS Lake Formation Developer Guide*.

See Also

- the section called “Logging and monitoring”

Cross-account resource ownership and billing

When a user in one AWS account (Account A) creates a new resource such as a database in a different account (Account B), that resource is then owned by Account B, the account where it was created. An administrator in Account B automatically gets full permissions to access the new resource, including reading, writing, and granting access permissions to a third account. The user in Account A can access the resource that they just created only if they have the appropriate permissions granted by Account B.

Storage costs and other costs that are directly associated with the new resource are billed to Account B, the resource owner. The cost of requests from the user who created the resource are billed to the requester's account, Account A.

For more information about AWS Glue billing and pricing, see How AWS Pricing Works.

Cross-account access limitations

AWS Glue cross-account access has the following limitations:

- Cross-account access to AWS Glue is not allowed if you created databases and tables using Amazon Athena orAmazon Redshift Spectrum prior to a region's support for AWS Glue and the resource owner account has not migrated the Amazon Athena data catalog to AWS Glue. You can find the current migration status using the GetCatalogImportStatus (get_catalog_import_status). For more details on how to migrate an Athena catalog to AWS Glue, see Upgrading to the AWS Glue Data Catalog step-by-step in the *Amazon Athena User Guide*.
- Cross-account access is *only* supported for Data Catalog resources, including databases, tables, user-defined functions, and connections.
- Cross-account access to the Data Catalog from Athena requires you to register the catalog as an Athena `DataCatalog` resource. For instructions, see Registering an AWS Glue Data Catalog from another account in the *Amazon Athena User Guide*.

Troubleshooting AWS Glue identity and access

Use the following information to help you diagnose and fix common issues that you might encounter when working with AWS Glue and IAM.

Topics

- I am not authorized to perform an action in AWS Glue
- I am not authorized to perform iam:PassRole
- I want to allow people outside of my AWS account to access my AWS Glue resources

I am not authorized to perform an action in AWS Glue

If you receive an error that you're not authorized to perform an action, your policies must be updated to allow you to perform the action.

The following example error occurs when the `mateojackson` IAM user tries to use the console to view details about a fictional *`my-example-widget`* resource but doesn't have the fictional `glue:`*`GetWidget`* permissions.

```
User: arn:aws:iam::123456789012:user/mateojackson is not authorized to perform:
 glue:GetWidget on resource: my-example-widget
```

In this case, the policy for the `mateojackson` user must be updated to allow access to the *`my-example-widget`* resource by using the `glue:`*`GetWidget`* action.

If you need help, contact your AWS administrator. Your administrator is the person who provided you with your sign-in credentials.

I am not authorized to perform iam:PassRole

If you receive an error that you're not authorized to perform the `iam:PassRole` action, your policies must be updated to allow you to pass a role to AWS Glue.

Some AWS services allow you to pass an existing role to that service instead of creating a new service role or service-linked role. To do this, you must have permissions to pass the role to the service.

The following example error occurs when an IAM user named `marymajor` tries to use the console to perform an action in AWS Glue. However, the action requires the service to have permissions that are granted by a service role. Mary does not have permissions to pass the role to the service.

```
User: arn:aws:iam::123456789012:user/marymajor is not authorized to perform:
 iam:PassRole
```

In this case, Mary's policies must be updated to allow her to perform the `iam:PassRole` action.

If you need help, contact your AWS administrator. Your administrator is the person who provided you with your sign-in credentials.

I want to allow people outside of my AWS account to access my AWS Glue resources

You can create a role that users in other accounts or people outside of your organization can use to access your resources. You can specify who is trusted to assume the role. For services that support resource-based policies or access control lists (ACLs), you can use those policies to grant people access to your resources.

To learn more, consult the following:

- To learn whether AWS Glue supports these features, see How AWS Glue works with IAM.
- To learn how to provide access to your resources across AWS accounts that you own, see Providing access to an IAM user in another AWS account that you own in the *IAM User Guide*.
- To learn how to provide access to your resources to third-party AWS accounts, see Providing access to AWS accounts owned by third parties in the *IAM User Guide*.

- To learn how to provide access through identity federation, see Providing access to externally authenticated users (identity federation) in the *IAM User Guide*.
- To learn the difference between using roles and resource-based policies for cross-account access, see Cross account resource access in IAM in the *IAM User Guide*.

AWS Lake Formation access control models

AWS Glue 5.0 supports two models for accessing data through AWS Lake Formation:

Topics

- Using AWS Glue with AWS Lake Formation for Full Table Access
- Using AWS Glue with AWS Lake Formation for fine-grained access control

Using AWS Glue with AWS Lake Formation for Full Table Access

Introduction to Full Table Access

AWS Glue 5.0 supports Full Table Access (FTA) control in Apache Spark based on your policies defined in AWS Lake Formation. This feature enables read and write operations from your AWS Glue Spark jobs on AWS Lake Formation registered tables when the job role has full table access. FTA is ideal for use cases that need to comply with security regulations at the table level and supports Spark capabilities including Resilient Distributed Datasets (RDDs), custom libraries, and User Defined Functions (UDFs) with AWS Lake Formation tables.

When a AWS Glue Spark job is configured for Full Table Access (FTA), AWS Lake Formation credentials are used to read/write Amazon S3 data for AWS Lake Formation registered tables, while the job's runtime role credentials will be used to read/write tables not registered with AWS Lake Formation. This capability enables Data Manipulation Language (DML) operations including CREATE, ALTER, DELETE, UPDATE, and MERGE INTO statements on Apache Hive and Iceberg tables.

> **Note**
>
> Review your requirements and determine if Fine-Grained Access Control (FGAC) or Full Table Access (FTA) suits your needs. Only one AWS Lake Formation permission method can be enabled for a given AWS Glue job. A job cannot simultaneously run Full Table Access (FTA) and Fine-Grained Access Control (FGAC) at the same time.

How Full-Table Access (FTA) works on AWS Glue

AWS Lake Formation offers two approaches for data access control: Fine-Grained Access Control (FGAC) and Full Table Access (FTA). FGAC provides enhanced security through column, row, and cell-level filtering, ideal for scenarios requiring granular permissions. FTA is ideal for straightforward access control scenarios where you need table- level permissions. It simplifies implementation by eliminating the need to enable fine-grained access mode, improves performance and reduces cost by avoiding the system driver and system executors, and supports both read and write operations (including CREATE, ALTER, DELETE, UPDATE, and MERGE INTO commands).

In AWS Glue 4.0, AWS Lake Formation based data access worked through GlueContext class, the utility class provided by AWS Glue. In AWS Glue 5.0, AWS Lake Formation based data access is available through native Spark SQL, Spark DataFrames, and continues to be supported through GlueContext class.

Implementing Full Table Access

Step 1: Enable Full Table Access in AWS Lake Formation

To use Full Table Access (FTA) mode, you need to allow third-party query engines to access data without the IAM session tag validation in AWS Lake Formation. To enable, follow the steps in Application integration for full table access .

Step 2: Setup IAM permissions for job runtime role

For read or write access to underlying data, in addition to AWS Lake Formation permissions, a job runtime role needs the `lakeformation:GetDataAccess` IAM permission. With this permission, AWS Lake Formation grants the request for temporary credentials to access the data.

The following is an example policy of how to provide IAM permissions to access a script in Amazon S3, uploading logs to Amazon S3, AWS Glue API permissions, and permission to access AWS Lake Formation.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
```

```
    "Sid": "ScriptAccess",
    "Effect": "Allow",
    "Action": [
      "s3:GetObject"
    ],
    "Resource": [
      "arn:aws:s3:::amzn-s3-demo-bucket/scripts/*"
    ]
  },
  {
    "Sid": "LoggingAccess",
    "Effect": "Allow",
    "Action": [
      "s3:PutObject"
    ],
    "Resource": [
      "arn:aws:s3:::amzn-s3-demo-bucket/logs/*"
    ]
  },
  {
    "Sid": "GlueCatalogAccess",
    "Effect": "Allow",
    "Action": [
      "glue:GetDatabase",
      "glue:GetDatabases",
      "glue:GetTable",
      "glue:GetTables",
      "glue:GetPartition",
      "glue:GetPartitions",
      "glue:CreateTable",
      "glue:UpdateTable"
    ],
    "Resource": [
      "arn:aws:glue:us-east-1:111122223333:catalog",
      "arn:aws:glue:us-east-1:111122223333:database/default",
      "arn:aws:glue:us-east-1:111122223333:table/default/*"
    ]
  },
  {
    "Sid": "LakeFormationAccess",
    "Effect": "Allow",
    "Action": [
      "lakeformation:GetDataAccess"
    ],
        "Resource": "*"
      }
    ]
  }
```

Step 2.1 Configure AWS Lake Formation permissions

AWS Glue Spark jobs that read data from Amazon S3 require AWS Lake Formation SELECT permission.

AWS Glue Spark jobs that write/delete data in Amazon S3 require AWS Lake Formation ALL permission.

AWS Glue Spark jobs that interact with AWS Glue Data catalog require DESCRIBE, ALTER, DROP permission as appropriate.

Step 3: Initialize a Spark session for Full Table Access using AWS Lake Formation

To access tables registered with AWS Lake Formation, the following configurations need to be set during Spark initialization to configure Spark to use AWS Lake Formation credentials.

To access tables registered with AWS Lake Formation, you need to explicitly configure your Spark session to use AWS Lake Formation credentials. Add the following configurations when initializing your Spark session:

```
from pyspark.sql import SparkSession

# Initialize Spark session with Lake Formation configurations
spark = SparkSession.builder \
    .appName("Lake Formation Full Table Access") \
    .config("spark.sql.catalog.glue_catalog",
 "org.apache.spark.sql.catalog.hive.GlueCatalog") \
    .config("spark.sql.catalog.glue_catalog.glue.lakeformation-enabled", "true") \
    .config("spark.sql.defaultCatalog", "glue_catalog") \
    .getOrCreate()
```

Key configurations:

- `spark.sql.catalog.glue_catalog`: Registers a catalog named "glue_catalog" that uses the GlueCatalog implementation

- `spark.sql.catalog.glue_catalog.glue.lakeformation-enabled`: Explicitly enables AWS Lake Formation integration for this catalog
- The catalog name ("glue_catalog" in this example) can be customized, but must be consistent in both configuration settings

Hive

```
--conf spark.hadoop.fs.s3.credentialsResolverClass=com.amazonaws.AWS
 Glue.accesscontrol.AWSLakeFormationCredentialResolver
--conf spark.hadoop.fs.s3.useDirectoryHeaderAsFolderObject=true
--conf spark.hadoop.fs.s3.folderObject.autoAction.disabled=true
--conf spark.sql.catalog.skipLocationValidationOnCreateTable.enabled=true
--conf spark.sql.catalog.createDirectoryAfterTable.enabled=true
--conf spark.sql.catalog.dropDirectoryBeforeTable.enabled=true
```

Iceberg

```
--conf spark.hadoop.fs.s3.credentialsResolverClass=com.amazonaws.AWS
 Glue.accesscontrol.AWSLakeFormationCredentialResolver
--conf spark.hadoop.fs.s3.useDirectoryHeaderAsFolderObject=true
--conf spark.hadoop.fs.s3.folderObject.autoAction.disabled=true
--conf spark.sql.catalog.skipLocationValidationOnCreateTable.enabled=true
--conf spark.sql.catalog.createDirectoryAfterTable.enabled=true
--conf spark.sql.catalog.dropDirectoryBeforeTable.enabled=true
--conf spark.sql.catalog.<catalog>.AWS Glue.lakeformation-enabled=true
```

- `spark.hadoop.fs.s3.credentialsResolverClass=com.amazonaws.AWS Glue.accesscontrol.AWSLakeFormationCredentialResolver`: Configure EMR Filesystem (EMRFS) to use AWS Lake Formation S3 credentials for AWS Lake Formation registered tables. If the table is not registered, use the job's runtime role credentials.
- `spark.hadoop.fs.s3.useDirectoryHeaderAsFolderObject=true` and `spark.hadoop.fs.s3.folderObject.autoAction.disabled=true`: Configure EMRFS to use content type header application/x-directory instead of $folder$ suffix when creating S3 folders. This is required when reading AWS Lake Formation tables, as AWS Lake Formation credentials do not allow reading table folders with $folder$ suffix.

- `spark.sql.catalog.skipLocationValidationOnCreateTable.enabled=true`: Configure Spark to skip validating the table location's emptiness before creation. This is necessary for AWS Lake Formation registered tables, as AWS Lake Formation credentials to verify the empty location are available only after AWS Glue Data Catalog table creation. Without this configuration, the job's runtime role credentials will validate the empty table location.
- `spark.sql.catalog.createDirectoryAfterTable.enabled=true`: Configure Spark to create the Amazon S3 folder after table creation in the Hive metastore. This is required for AWS Lake Formation registered tables, as AWS Lake Formation credentials to create the Amazon S3 folder are available only after AWS Glue Data Catalog table creation.
- `spark.sql.catalog.dropDirectoryBeforeTable.enabled=true`: Configure Spark to drop the Amazon S3 folder before table deletion in the Hive metastore. This is necessary for AWS Lake Formation registered tables, as AWS Lake Formation credentials to drop the S3 folder are not available after table deletion from the AWS Glue Data Catalog.
- `spark.sql.catalog.<catalog>.AWS Glue.lakeformation-enabled=true`: Configure Iceberg catalog to use AWS Lake Formation Amazon S3 credentials for AWS Lake Formation registered tables. If the table is not registered, use default environment credentials.

Usage Patterns

Using FTA with DataFrames

For users familiar with Spark, DataFrames can be used with AWS Lake Formation Full Table Access.

AWS Glue 5.0 adds native Spark support for Lake Formation Full Table Access, simplifying how you work with protected tables. This feature enables AWS Glue 5.0 AWS Glue Spark jobs to directly read and write data when full table access is granted, removing limitations that previously restricted certain Extract, Transform, and Load (ETL) operations. You can now leverage advanced Spark capabilities including Resilient Distributed Datasets (RDDs), custom libraries, and User Defined Functions (UDFs) with AWS Lake Formation tables.

Native Spark FTA in AWS Glue 5.0

AWS Glue 5.0 supports full-table access (FTA) control in Apache Spark based on your policies defined in AWS Lake Formation. This level of control is ideal for use cases that need to comply with security regulations at the table level.

Apache Iceberg Table Example

```
from pyspark.sql import SparkSession

catalog_name = "spark_catalog"
aws_region = "us-east-1"
aws_account_id = "123456789012"
warehouse_path = "s3://amzn-s3-demo-bucket/warehouse/"

spark = SparkSession.builder \

 .config("spark.sql.extensions","org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensi
 \
    .config(f"spark.sql.catalog.{catalog_name}",
 "org.apache.iceberg.spark.SparkSessionCatalog") \
    .config(f"spark.sql.catalog.{catalog_name}.warehouse", f"{warehouse_path}") \
    .config(f"spark.sql.catalog.{catalog_name}.client.region",f"{aws_region}") \
    .config(f"spark.sql.catalog.{catalog_name}.glue.account-id",f"{aws_account_id}") \
    .config(f"spark.sql.catalog.{catalog_name}.glue.lakeformation-enabled","true") \
    .config(f"spark.sql.catalog.dropDirectoryBeforeTable.enabled", "true") \
    .config(f"spark.sql.catalog.{catalog_name}.catalog-impl",
 "org.apache.iceberg.aws.glue.GlueCatalog") \
    .config(f"spark.sql.catalog.{catalog_name}.io-impl",
 "org.apache.iceberg.aws.s3.S3FileIO") \
    .config("spark.sql.defaultCatalog", catalog_name) \  # Add this line
    .getOrCreate()

database_name = "your_database"
table_name = "your_table"

df = spark.sql(f"select * from {database_name}.{table_name}")
df.show()
```

Required IAM Permissions

Your AWS Glue job execution role must have:

```
{
    "Action": "lakeformation:GetDataAccess",
    "Resource": "*",
    "Effect": "Allow"
}
```

Plus appropriate S3 access permissions for your data locations.

Lake Formation Configuration

Before using native Spark FTA in AWS Glue 5.0:

1. Allow third-party query engines to access data without IAM session tag validation in AWS Lake Formation
2. Grant appropriate table permissions to your AWS Glue job execution role through AWS Lake Formation console
3. Configure your Spark session with the required parameters shown in the example above

Using FTA with DynamicFrames

AWS Glue's native DynamicFrames can be used with AWS Lake Formation Full Table Access for optimized ETL operations. Full Table Access (FTA) provides a security model that grants permissions at the table level, allowing for faster data processing compared to Fine-Grained Access Control (FGAC) since it bypasses the overhead of row and column-level permission checks. This approach is useful when you need to process entire tables and table-level permissions meet your security requirements.

In AWS Glue 4.0, DynamicFrames with FTA required specific GlueContext configuration. While existing AWS Glue 4.0 DynamicFrame code with FTA will continue to work in AWS Glue 5.0, the newer version also offers native Spark FTA support with greater flexibility. For new development, consider using the native Spark approach described in the DataFrames section, especially if you need additional capabilities such as Resilient Distributed Datasets (RDDs), custom libraries, and User Defined Functions (UDFs) with AWS Lake Formation tables.

Required Permissions

The IAM role executing your Glue job must have:

- `lakeformation:GetDataAccess` permission
- Appropriate Lake Formation table permissions granted through the Lake Formation console

Example DynamicFrame Implementation in AWS Glue 5.0

```
from awsglue.context import GlueContext
from pyspark.context import SparkContext

# Initialize Glue context
sc = SparkContext()
glueContext = GlueContext(sc)

# Configure catalog for Iceberg tables
catalog_name = "glue_catalog"
aws_region = "us-east-1"
aws_account_id = "123456789012"
warehouse_path = "s3://amzn-s3-demo-bucket/warehouse/"

spark = glueContext.spark_session
spark.conf.set(f"spark.sql.catalog.{catalog_name}",
 "org.apache.iceberg.spark.SparkCatalog")
spark.conf.set(f"spark.sql.catalog.{catalog_name}.warehouse", f"{warehouse_path}")
spark.conf.set(f"spark.sql.catalog.{catalog_name}.catalog-impl",
 "org.apache.iceberg.aws.glue.GlueCatalog")
spark.conf.set(f"spark.sql.catalog.{catalog_name}.io-impl",
 "org.apache.iceberg.aws.s3.S3FileIO")
spark.conf.set(f"spark.sql.catalog.{catalog_name}.glue.lakeformation-enabled","true")
spark.conf.set(f"spark.sql.catalog.{catalog_name}.client.region",f"{aws_region}")
spark.conf.set(f"spark.sql.catalog.{catalog_name}.glue.id", f"{aws_account_id}")

# Read Lake Formation-protected table with DynamicFrame
df = glueContext.create_data_frame.from_catalog(
    database="your_database",
    table_name="your_table"
)
```

Additional Configuration

Configure full table access mode in AWS Glue Studio notebooks

To access AWS Lake Formation registered tables from interactive Spark sessions in AWS Glue Studio notebooks, you must use compatibility permission mode. Use the `%%configure` magic command to set up your Spark configuration before starting your interactive session. This configuration must be the first command in your notebook, as it cannot be applied after the session has started. Choose the configuration based on your table type:

For Hive tables

```
%%configure
```

```
--conf
 spark.hadoop.fs.s3.credentialsResolverClass=com.amazonaws.glue.accesscontrol.AWSLakeFormationC
--conf spark.hadoop.fs.s3.useDirectoryHeaderAsFolderObject=true
--conf spark.hadoop.fs.s3.folderObject.autoAction.disabled=true
--conf spark.sql.catalog.skipLocationValidationOnCreateTable.enabled=true
--conf spark.sql.catalog.createDirectoryAfterTable.enabled=true
--conf spark.sql.catalog.dropDirectoryBeforeTable.enabled=true
```

For Iceberg tables

```
%%configure
--conf
 spark.hadoop.fs.s3.credentialsResolverClass=com.amazonaws.glue.accesscontrol.AWSLakeFormationC
--conf spark.hadoop.fs.s3.useDirectoryHeaderAsFolderObject=true
--conf spark.hadoop.fs.s3.folderObject.autoAction.disabled=true
--conf spark.sql.catalog.skipLocationValidationOnCreateTable.enabled=true
--conf spark.sql.catalog.createDirectoryAfterTable.enabled=true
--conf spark.sql.catalog.dropDirectoryBeforeTable.enabled=true
--conf spark.sql.catalog.glue_catalog.glue.lakeformation-enabled=true
--conf spark.sql.catalog.glue_catalog.warehouse=s3://example-s3-bucket_DATA_LOCATION
--conf spark.sql.catalog.glue_catalog.catalog-
impl=org.apache.iceberg.aws.glue.GlueCatalog
--conf spark.sql.catalog.glue_catalog.io-impl=org.apache.iceberg.aws.s3.S3FileIO
--conf spark.sql.catalog.glue_catalog.glue.account-id=ACCOUNT_ID
--conf spark.sql.catalog.glue_catalog.glue.region=REGION
```

Replace the placeholders:

- S3_DATA_LOCATION: *`s3://amzn-s3-demo-bucket`*
- REGION: *`AWS Region (e.g., us-east-1)`*
- ACCOUNT_ID: *`Your AWS Account ID`*

> **Note**
>
> You must set these configurations before executing any Spark operations in your notebook.

Supported Operations

These operations will use AWS Lake Formation credentials to access the table data.

> **Note**
>
> On enabling AWS Lake Formation:
>
> - For FTA: Enable the Spark configuration `spark.sql.catalog.{catalog_name}.glue.lakeformation-enabled`

- CREATE TABLE
- ALTER TABLE
- INSERT INTO
- INSERT OVERWRITE
- SELECT
- UPDATE
- MERGE INTO
- DELETE FROM
- ANALYZE TABLE
- REPAIR TABLE
- DROP TABLE
- Spark datasource queries
- Spark datasource writes

> **Note**
>
> Operations not listed above will continue to use IAM permissions to access table data.

Migrating from AWS Glue 4.0 to AWS Glue 5.0 FTA

When migrating from AWS Glue 4.0 GlueContext FTA to AWS Glue 5.0 native Spark FTA:

1. Allow third-party query engines to access data without the IAM session tag validation in AWS Lake Formation. Follow Step 1: Enable Full Table Access in AWS Lake Formation.
2. You do not need to change the job runtime role. However, verify that the AWS Glue job execution role has lakeformation:GetDataAccess IAM permission.

3. Modify spark session configurations in the script. Ensure the following spark configurations are present:

```
--conf spark.sql.catalog.spark_catalog=org.apache.iceberg.spark.SparkSessionCatalog
--conf spark.sql.catalog.spark_catalog.warehouse=s3://<bucket-name>/warehouse/
--conf spark.sql.catalog.spark_catalog.client.region=<REGION>
--conf spark.sql.catalog.spark_catalog.glue.account-id=ACCOUNT_ID
--conf spark.sql.catalog.spark_catalog.glue.lakeformation-enabled=true
--conf spark.sql.catalog.dropDirectoryBeforeTable.enabled=true
```

4. Update script such that GlueContext DataFrames are changed to native spark DataFrames.
5. Update your AWS Glue job to use AWS Glue 5.0

Considerations and Limitations

- If a Hive table is created using a job that doesn't have full table access enabled, and no records are inserted, subsequent reads or writes from a job with full table access will fail. This is because AWS Glue Spark without full table access adds the $folder$ suffix to the table folder name. To resolve this, you can either:
  - Insert at least one row into the table from a job that does not have FTA enabled.
  - Configure the job that does not have FTA enabled to not use $folder$ suffix in folder name in S3. This can be achieved by setting Spark configuration `spark.hadoop.fs.s3.useDirectoryHeaderAsFolderObject=true`.
  - Create an Amazon S3 folder at the table location `s3://path/to/table/table_name` using the Amazon S3 console or Amazon S3 CLI.
- Full Table Access works exclusively with EMR Filesystem (EMRFS). S3A filesystem is not compatible.
- Full Table Access is supported for Hive and Iceberg tables. Support for Hudi and Delta tables has not yet been added.
- Jobs referencing tables with AWS Lake Formation Fine-Grained Access Control (FGAC) rules or AWS Glue Data Catalog Views will fail. To query a table with an FGAC rules or a AWS Glue Data Catalog View, you need to use the FGAC mode. You can enable FGAC mode by following the steps outlined in the AWS documentation: Using AWS Glue with AWS Lake Formation for fine-grained access control.
- Full table access does not support Spark Streaming.
- Cannot be used simultaneously with FGAC.

Using AWS Glue with AWS Lake Formation for fine-grained access control

Overview

With AWS Glue version 5.0 and higher, you can leverage AWS Lake Formation to apply fine-grained access controls on Data Catalog tables that are backed by S3. This capability lets you configure table, row, column, and cell level access controls for read queries within your AWS Glue for Apache Spark jobs. See the following sections to learn more about Lake Formation and how to use it with AWS Glue.

`GlueContext`-based table-level access control with AWS Lake Formation permissions supported in Glue 4.0 or before is not supported in Glue 5.0. Use the new Spark native fine-grained access control (FGAC) in Glue 5.0. Note the following details:

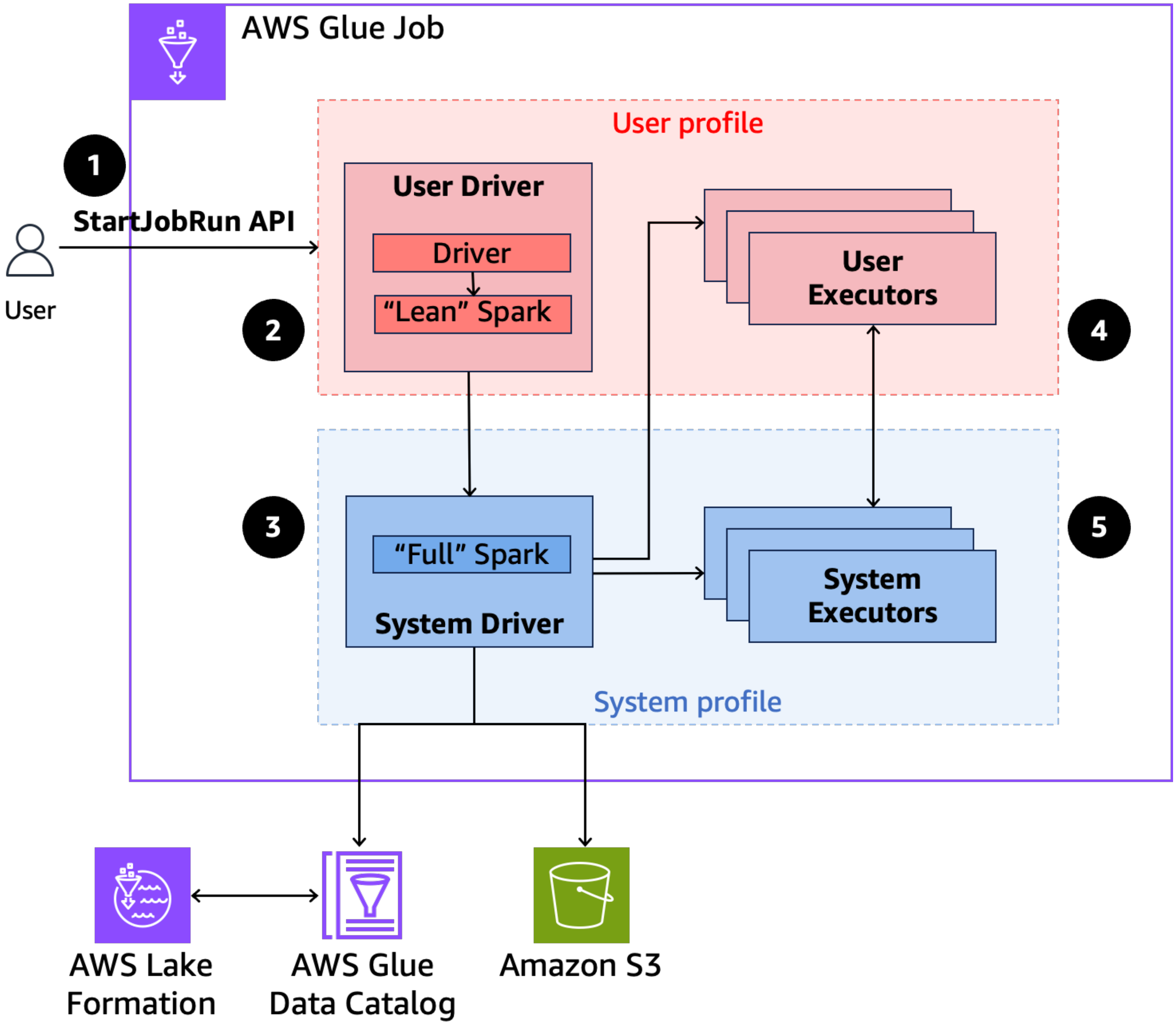
- If you need fine grained access control (FGAC) for row/column/cell access control, you will need to migrate from `GlueContext`/Glue DynamicFrame in Glue 4.0 and prior to Spark dataframe in Glue 5.0. For examples, see Migrating from GlueContext/Glue DynamicFrame to Spark DataFrame
- If you need Full Table Access control (FTA), you can leverage FTA with DynamicFrames in AWS Glue 5.0. You can also migrate to native Spark approach for additional capabilities such as Resilient Distributed Datasets (RDDs), custom libraries, and User Defined Functions (UDFs) with AWS Lake Formation tables. For examples, see Migrating from AWS Glue 4.0 to AWS Glue 5.0.
- If you don't need FGAC, then no migration to Spark dataframe is necessary and `GlueContext` features like job bookmarks, push down predicates will continue to work.
- Jobs with FGAC require a minimum of 4 workers: one user driver, one system driver, one system executor, and one standby user executor.

Using AWS Glue with AWS Lake Formation incurs additional charges.

How AWS Glue works with AWS Lake Formation

Using AWS Glue with Lake Formation lets you enforce a layer of permissions on each Spark job to apply Lake Formation permissions control when AWS Glue executes jobs. AWS Glue uses Spark resource profiles to create two profiles to effectively execute jobs. The user profile executes user-supplied code, while the system profile enforces Lake Formation policies. For more information, see What is AWS Lake Formation and Considerations and limitations.

The following is a high-level overview of how AWS Glue gets access to data protected by Lake Formation security policies.


AWS Glue Job
User profile
User Driver
Driver
"Lean" Spark
User Executors
1
StartJobRun API
User
2
3
4
5
"Full" Spark
System Driver
System Executors
System profile
AWS Lake Formation
AWS Glue Data Catalog
Amazon S3

1. A user calls the `StartJobRun` API on an AWS Lake Formation-enabled AWS Glue job.
2. AWS Glue sends the job to a user driver and runs the job in the user profile. The user driver runs a lean version of Spark that has no ability to launch tasks, request executors, access S3 or the Glue Catalog. It builds a job plan.
3. AWS Glue sets up a second driver called the system driver and runs it in the system profile (with a privileged identity). AWS Glue sets up an encrypted TLS channel between the two drivers for communication. The user driver uses the channel to send the job plans to the

system driver. The system driver does not run user-submitted code. It runs full Spark and communicates with S3, and the Data Catalog for data access. It request executors and compiles the Job Plan into a sequence of execution stages.

4. AWS Glue then runs the stages on executors with the user driver or system driver. User code in any stage is run exclusively on user profile executors.
5. Stages that read data from Data Catalog tables protected by AWS Lake Formation or those that apply security filters are delegated to system executors.

Minimum worker requirement

A Lake Formation-enabled job in AWS Glue requires a minimum of 4 workers: one user driver, one system driver, one system executor, and one standby User Executor. This is up from the minimum of 2 workers required for standard AWS Glue jobs.

A Lake Formation-enabled job in AWS Glue utilizes two Spark drivers—one for the system profile and another for the user profile. Similarly, the executors are also divided into two profiles:

- System executors: handle tasks where Lake Formation data filters are applied.
- User executors: are requested by the system driver as needed.

As Spark jobs are lazy in nature, AWS Glue reserves 10% of the total workers (minimum of 1), after deducting the two drivers, for user executors.

All Lake Formation-enabled jobs have auto-scaling enabled, meaning the user executors will only start when needed.

For an example configuration, see Considerations and limitations.

Job runtime role IAM permissions

Lake Formation permissions control access to AWS Glue Data Catalog resources, Amazon S3 locations, and the underlying data at those locations. IAM permissions control access to the Lake Formation and AWS Glue APIs and resources. Although you might have the Lake Formation permission to access a table in the Data Catalog (SELECT), your operation fails if you don't have the IAM permission on the `glue:Get*` API operation.

The following is an example policy of how to provide IAM permissions to access a script in S3, uploading logs to S3, AWS Glue API permissions, and permission to access Lake Formation.

JSON

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "ScriptAccess",
      "Effect": "Allow",
      "Action": [
        "s3:GetObject",
        "s3:ListBucket"
      ],
      "Resource": [
        "arn:aws:s3:::*.amzn-s3-demo-bucket/scripts",
        "arn:aws:s3:::*.amzn-s3-demo-bucket/*"
      ]
    },
    {
      "Sid": "LoggingAccess",
      "Effect": "Allow",
      "Action": [
        "s3:PutObject"
      ],
      "Resource": [
        "arn:aws:s3:::amzn-s3-demo-bucket/logs/*"
      ]
    },
    {
      "Sid": "GlueCatalogAccess",
      "Effect": "Allow",
      "Action": [
        "glue:Get*",
        "glue:Create*",
        "glue:Update*"
      ],
      "Resource": [
        "*"
      ]
    },
    {
      "Sid": "LakeFormationAccess",
      "Effect": "Allow",
      "Action": [
          "lakeformation:GetDataAccess"
        ],
        "Resource": [
          "*"
        ]
      }
    ]
  }
```

Setting up Lake Formation permissions for job runtime role

First, register the location of your Hive table with Lake Formation. Then create permissions for your job runtime role on your desired table. For more details about Lake Formation, see [What is AWS Lake Formation?](#) in the *AWS Lake Formation Developer Guide*.

After you set up the Lake Formation permissions, you can submit Spark jobs on AWS Glue.

Submitting a job run

After you finish setting up the Lake Formation grants, you can submit Spark jobs on AWS Glue. To run Iceberg jobs, you must provide the following Spark configurations. To configure through Glue job parameters, put the following parameter:

- Key:

  ```
  --conf
  ```

- Value:

  ```
  spark.sql.catalog.spark_catalog=org.apache.iceberg.spark.SparkSessionCatalog
         --conf spark.sql.catalog.spark_catalog.warehouse=<S3_DATA_LOCATION>
         --conf spark.sql.catalog.spark_catalog.glue.account-id=<ACCOUNT_ID>
         --conf spark.sql.catalog.spark_catalog.client.region=<REGION>
         --conf spark.sql.catalog.spark_catalog.glue.endpoint=https://
  glue.<REGION>.amazonaws.com
  ```

Using an Interactive Session

After you finish setting up the AWS Lake Formation grants, you can use Interactive Sessions on AWS Glue. You must provide the following Spark configurations via the `%%configure` magic prior to executing code.

```
%%configure
{
    "--enable-lakeformation-fine-grained-access": "true",
    "--conf":
 "spark.sql.catalog.spark_catalog=org.apache.iceberg.spark.SparkSessionCatalog
 --conf spark.sql.catalog.spark_catalog.warehouse=<S3_DATA_LOCATION> --conf
 spark.sql.catalog.spark_catalog.catalog-impl=org.apache.iceberg.aws.glue.GlueCatalog
 --conf spark.sql.catalog.spark_catalog.io-impl=org.apache.iceberg.aws.s3.S3FileIO --
conf
 spark.sql.extensions=org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions
 --conf spark.sql.catalog.spark_catalog.client.region=<REGION> --conf
 spark.sql.catalog.spark_catalog.glue.account-id=<ACCOUNT_ID> --conf
 spark.sql.catalog.spark_catalog.glue.endpoint=https://glue.<REGION>.amazonaws.com"
}
```

FGAC for AWS Glue 5.0 Notebook or interactive sessions

To enable Fine-Grained Access Control (FGAC) in AWS Glue you must specify the Spark confs required for Lake Formation as part of the %%configure magic before you create first cell.

Specifying it later using the calls `SparkSession.builder().conf("").get()` or `SparkSession.builder().conf("").create()` will not be enough. This is a change from the AWS Glue 4.0 behavior.

Open-table format support

AWS Glue version 5.0 or later includes support for fine-grained access control based on Lake Formation. AWS Glue supports Hive and Iceberg table types. The following table describes all of the supported operations.

Operations	Hive	Iceberg
DDL commands	With IAM role permissions only	With IAM role permissions only

Operations	Hive	Iceberg
Incremental queries	Not applicable	Fully supported
Time travel queries	Not applicable to this table format	Fully supported
Metadata tables	Not applicable to this table format	Supported, but certain tables are hidden. See [considerations and limitations](#) for more information.
`DML INSERT`	With IAM permissions only	With IAM permissions only
DML UPDATE	Not applicable to this table format	With IAM permissions only
`DML DELETE`	Not applicable to this table format	With IAM permissions only
Read operations	Fully supported	Fully supported
Stored procedures	Not applicable	Supported with the exceptions of `register_table` and `migrate`. See [considerations and limitations](#) for more information.

Migrating from GlueContext/Glue DynamicFrame to Spark DataFrame

The following are Python and Scala examples of migrating `GlueContext`/Glue `DynamicFrame` in Glue 4.0 to Spark `DataFrame` in Glue 5.0.

Python

Before:

```
escaped_table_name= '`<dbname>`.`<table_name>`'

additional_options = {
  "query": f'select * from {escaped_table_name} WHERE column1 = 1 AND column7 = 7'
}

# DynamicFrame example
dataset = glueContext.create_data_frame_from_catalog(
    database="<dbname>",
    table_name=escaped_table_name,
    additional_options=additional_options)
```

After:

```
table_identifier= '`<catalogname>`.`<dbname>`.`<table_name>`"' #catalogname is optional

# DataFrame example
dataset = spark.sql(f'select * from {table_identifier} WHERE column1 = 1 AND column7 =
 7')
```

Scala

Before:

```
val escapedTableName = "`<dbname>`.`<table_name>`"

val additionalOptions = JsonOptions(Map(
    "query" -> s"select * from $escapedTableName WHERE column1 = 1 AND column7 = 7"
    )
)

# DynamicFrame example
val datasource0 = glueContext.getCatalogSource(
    database="<dbname>",
    tableName=escapedTableName,
    additionalOptions=additionalOptions).getDataFrame()
```

After:

```
val tableIdentifier = "`<catalogname>`.`<dbname>`.`<table_name>`" //catalogname is
 optional

# DataFrame example
val datasource0 = spark.sql(s"select * from $tableIdentifier WHERE column1 = 1 AND
 column7 = 7")
```

Considerations and limitations

Consider the following considerations and limitations when you use Lake Formation with AWS Glue.

AWS Glue with Lake Formation is available in all supported Regions except AWS GovCloud (US-East) and AWS GovCloud (US-West).

- AWS Glue supports fine-grained access control via Lake Formation only for Apache Hive and Apache Iceberg tables. Apache Hive formats include Parquet, ORC, and CSV.
- You can only use Lake Formation with Spark jobs.
- AWS Glue with Lake Formation only supports a single Spark session throughout a job.
- When Lake Formation is enabled,AWS Glue requires a greater number of workers because it requires one system driver, system executors, one user driver, and optionally user executors (required when your job has UDFs or `spark.createDataFrame`).
- AWS Glue with Lake Formation only supports cross-account table queries shared through resource links. The resource-link needs to be named identically to the source account's resource.
- To enable fine-grained access control for AWS Glue jobs, pass the `--enable-lakeformation-fine-grained-access` job parameter.
- You can configure your AWS Glue jobs to work with the AWS Glue multi-catalog hierarchy. For information on the configuration parameters to use with the AWS Glue `StartJobRun` API, see Working with AWS Glue multi-catalog hierarchy on EMR Serverless.
- The following aren't supported:
  - Resilient distributed datasets (RDD)
  - Spark streaming
  - Write with Lake Formation granted permissions
  - Access control for nested columns
- AWS Glue blocks functionalities that might undermine the complete isolation of system driver, including the following:
  - UDTs, HiveUDFs, and any user-defined function that involves custom classes
  - Custom data sources
  - Supply of additional jars for Spark extension, connector, or metastore
  - `ANALYZE TABLE` command

- To enforce access controls, `EXPLAIN PLAN` and DDL operations such as `DESCRIBE TABLE` don't expose restricted information.
- AWS Glue restricts access to system driver Spark logs on Lake Formation-enabled applications. Since the system driver runs with more access, events and logs that the system driver generates can include sensitive information. To prevent unauthorized users or code from accessing this sensitive data, AWS Glue disabled access to system driver logs. For troubleshooting, contact AWS support.
- If you registered a table location with Lake Formation, the data access path goes through the Lake Formation stored credentials regardless of the IAM permission for the AWS Glue job runtime role. If you misconfigure the role registered with table location, jobs submitted that use the role with S3 IAM permission to the table location will fail.
- Writing to a Lake Formation table uses IAM permission rather than Lake Formation granted permissions. If your job runtime role has the necessary S3 permissions, you can use it to run write operations.

The following are considerations and limitations when using Apache Iceberg:

- You can only use Apache Iceberg with session catalog and not arbitrarily named catalogs.
- Iceberg tables that are registered in Lake Formation only support the metadata tables `history`, `metadata_log_entries`, `snapshots`, `files`, `manifests`, and `refs`. AWS Glue hides the columns that might have sensitive data, such as `partitions`, `path`, and `summaries`. This limitation doesn't apply to Iceberg tables that aren't registered in Lake Formation.
- Tables that you don't register in Lake Formation support all Iceberg stored procedures. The `register_table` and `migrate` procedures aren't supported for any tables.
- We recommend that you use Iceberg DataFrameWriterV2 instead of V1.

Example worker allocation

For a job configured with the following parameters:

```
--enable-lakeformation-fine-grained-access=true
--number-of-workers=20
```

The worker allocation would be:

- One worker for the user driver.

- One worker for the system driver.
- 10% of the remaining 18 workers (that is, 2 workers) reserved for the user executors.
- Up to 16 workers allocated for system executors.

With auto-scaling enabled, the user executors can utilize any of the unallocated capacity from the system executors if needed.

Controlling user executor allocation

You can adjust the reservation percentage for user executors using the following configuration:

```
--conf spark.dynamicAllocation.maxExecutorsRatio=<value between 0 and 1>
```

This configuration allows fine-tuned control over how many user executors are reserved relative to the total available capacity.

Troubleshooting

See the following sections for troubleshooting solutions.

Logging

AWS Glue uses Spark resources profiles to split job execution. AWS Glue uses the user profile to run the code you supplied, while the system profile enforces Lake Formation policies. You can access the logs for the tasks ran as the user profile.

Live UI and Spark History Server

The Live UI and the Spark History Server have all Spark events generated from the user profile and redacted events generated from the system driver.

You can see all of the tasks from both the user and system drivers in the **Executors** tab. However, log links are available only for the user profile. Also, some information is redacted from Live UI, such as the number of output records.

Job failed with insufficient Lake Formation permissions

Make sure that your job runtime role has the permissions to run SELECT and DESCRIBE on the table that you are accessing.

Job with RDD execution failed

AWS Glue currently doesn't support resilient distributed dataset (RDD) operations on Lake Formation-enabled jobs.

Unable to access data files in Amazon S3

Make sure you have registered the location of the data lake in Lake Formation.

Security validation exception

AWS Glue detected a security validation error. Contact AWS support for assistance.

Sharing AWS Glue Data Catalog and tables across accounts

You can share databases and tables across accounts and still use Lake Formation. For more information, see Cross-account data sharing in Lake Formation and How do I share AWS Glue Data Catalog and tables cross-account using ?.

The following table describes summary of how to choose between Fine-grained access control (FGAC) and Full table access (FTA) for your workload.

Feature	Fine-grained access control (FGAC)	Full table access (FTA)
Access Level	Column/row level	Full table
Use Case	Queries and ETL with limited permissions	ETL
Performance Impact	Requires system/user space transitions for access control evaluation, adding latency	Optimized performance

Using Amazon S3 Access Grants with AWS Glue

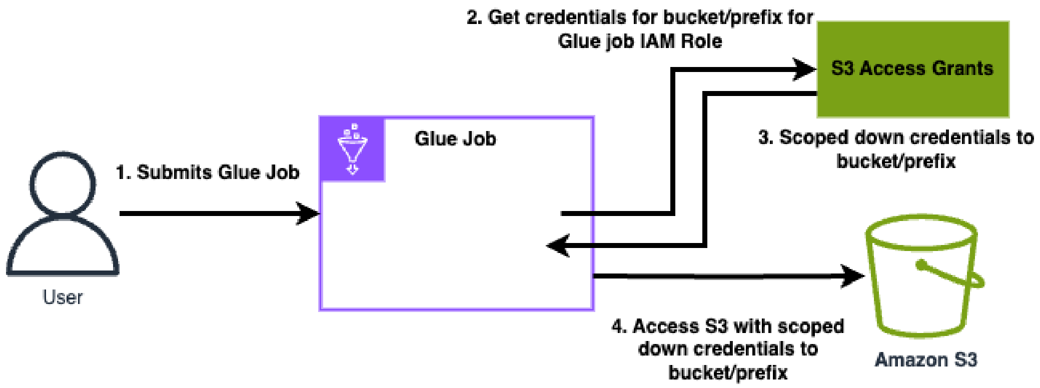
With Glue version 5.0, Amazon S3 Access Grants provide a scalable access control solution that you can use to augment access to your Amazon S3 data from AWS Glue. If you have a complex or large permission configuration for your S3 data, you can use S3 Access Grants to scale S3 data permissions for users and roles.

Use S3 Access Grants to augment access to Amazon S3 data beyond the permissions that are granted by the runtime role or the IAM roles that are attached to the identities with access to your AWS Glue job. For more information, see [Managing access with S3 Access Grants]() in the *Amazon S3 User Guide*.

How AWS Glue works with S3 Access Grants

AWS Glue versions 5.0 and higher provide a native integration with S3 Access Grants. You can enable S3 Access Grants on AWS Glue and run Spark jobs. When a Spark job makes a request for S3 data, Amazon S3 provides temporary credentials that are scoped to the specific bucket, prefix, or object.

The following is a high-level overview of how AWS Glue gets access to data that S3 Access Grants manages access to.


2. Get credentials for bucket/prefix for Glue job IAM Role
S3 Access Grants
Glue Job
1. Submits Glue Job
3. Scoped down credentials to bucket/prefix
User
4. Access S3 with scoped down credentials to bucket/prefix
Amazon S3

1. A user submits an AWS Glue Spark job that uses data stored in Amazon S3.
2. AWS Glue makes a request for S3 Access Grants to vend temporary credentials for the user that give access to the bucket, prefix, or object.
3. AWS Glue returns temporary credentials in the form of an AWS Security Token Service (STS) token for the user. The token is scoped to access the S3 bucket, prefix, or object.
4. AWS Glue uses the STS token to retrieve data from S3.
5. AWS Glue receives the data from S3 and returns the results to the user.

S3 Access Grants considerations with AWS Glue

Take note of the following behaviors and limitations when you use S3 Access Grants with AWS Glue.

Feature support

- S3 Access Grants is supported with AWS Glue versions 5.0 and higher.
- Spark is the only supported job type when you use S3 Access Grants with AWS Glue.
- Delta Lake and Hudi are the only supported open-table formats when you use S3 Access Grants with AWS Glue.
- The following capabilities are not supported for use with S3 Access Grants:
  - Apache Iceberg tables
  - AWS CLI requests to Amazon S3 that use IAM roles
  - S3 access through the open-source S3A protocol

Behavioral considerations

- AWS Glue provides a credentials cache to ensure that a user doesn't need to make repeated requests for the same credentials within a Spark job. Therefore, AWS Glue always requests the default-level privilege when it requests credentials. For more information, see Request access to S3 data in the *Amazon S3 User Guide*.

Set up S3 Access Grants with AWS Glue

Prerequisites

The caller or admin has created an S3 Access Grants instance.

Set up AWS Glue policies and job configuration

To set up S3 Access Grants with AWS Glue you must configure trust and IAM policies, and pass the configuration through job parameters.

1. Configure the following minimal trust and IAM policies on the role used for grants (the AWS Glue role that runs sessions or jobs).

   Trust policy:

   ```
   {
                "Sid": "Stmt1234567891011",
                "Action": [
                    "sts:AssumeRole",
   ```

```
                "sts:SetSourceIdentity",
                "sts:SetContext"
            ],
            "Effect": "Allow",
            "Principal": {
                "Service": "access-grants.s3.amazonaws.com"
            },
            "Condition": {
                "StringEquals": {
                    "aws:SourceAccount": "123456789012",
                    "aws:SourceArn": "arn:aws:s3:<region>:123456789012:access-grants/
default"
                }
            }
        }
```

IAM policy:

```
{
            "Sid": "S3Grants",
            "Effect": "Allow",
            "Action": [
                "s3:GetDataAccess",
                "s3:GetAccessGrantsInstanceForPrefix"
            ],
            "Resource": "arn:aws:s3:<region>:123456789012:access-grants/default"
        },
        {
            "Sid": "BucketLevelReadPermissions",
            "Effect": "Allow",
            "Action": [
                "s3:ListBucket"
            ],
            "Resource": [
                "arn:aws:s3:::*"
            ],
            "Condition": {
                "StringEquals": {
                    "aws:ResourceAccount": "123456789012"
                },
                "ArnEquals": {
                    "s3:AccessGrantsInstanceArn": [
                        "arn:aws:s3:<region>:123456789012:access-grants/default"
                    ]
                }
            }
        },
        {
            "Sid": "ObjectLevelReadPermissions",
            "Effect": "Allow",
            "Action": [
                "s3:GetObject",
                "s3:GetObjectVersion",
                "s3:GetObjectAcl",
                "s3:GetObjectVersionAcl",
                "s3:ListMultipartUploadParts"
            ],
            "Resource": [
                "arn:aws:s3:::*"
            ],
            "Condition": {
                "StringEquals": {
                    "aws:ResourceAccount": "123456789012"
                },
                "ArnEquals": {
                    "s3:AccessGrantsInstanceArn": [
                        "arn:aws:s3:<region>:123456789012:access-grants/default"
                    ]
                }
            }
        }
```

2. In your AWS Glue job, pass the following Spark configuration either through AWS Glue job parameters or `SparkConf`.

```
--conf spark.hadoop.fs.s3.s3AccessGrants.enabled=true \
--conf spark.hadoop.fs.s3.s3AccessGrants.fallbackToIAM=false
```

Trusted Identity Propagation with AWS Glue ETL

With IAM Identity Center, you can connect to identity providers (IdPs) and centrally manage access for users and groups across AWS analytics services. You can integrate identity providers such as Okta, Ping, and Microsoft Entra ID (formerly Azure Active Directory) with IAM Identity Center for users in your organization to access data using a single-sign on experience. IAM Identity Center also supports connecting additional third-party identity providers.

With AWS Glue 5.0 and higher, you can propagate user-identities from IAM Identity Center to AWS Glue interactive sessions. AWS Glue Interactive Sessions will further propagate supplied identity to downstream services such as Amazon S3 Access Grants, AWS Lake Formation, and Amazon Redshift, enabling secure data access via user identity in these downstream services.

Overview

Identity Center is the recommended approach for workforce authentication and authorization on AWS for organizations of any size and type. With Identity Center, you can create and manage user identities in AWS, or connect your existing identity source, including Microsoft Active Directory, Okta, Ping Identity, JumpCloud, Google Workspace, and Microsoft Entra ID (formerly Azure AD).

Trusted identity propagation is an IAM Identity Center feature that administrators of connected AWS services can use to grant and audit access to service data. Access to this data is based on user attributes such as group associations. Setting up trusted identity propagation requires collaboration between the administrators of connected AWS services and the IAM Identity Center administrators.

Features and benefits

The AWS Glue interactive sessions integration with IAM Identity Center Trusted identity propagation provides the following benefits:

- The ability to enforce table-level authorization and fine grained access control with Identity Center identities on Lake Formation managed AWS Glue data catalog tables.
- The ability to enforce authorization with Identity Center identities on Amazon Redshift clusters.
- Enables end to end tracking of user actions for auditing.
- The ability to enforce Amazon S3 prefix-level authorization with Identity Center identities on Amazon S3 Access Grants-managed Amazon S3 prefixes.

Use cases

Interactive Data Exploration and Analysis

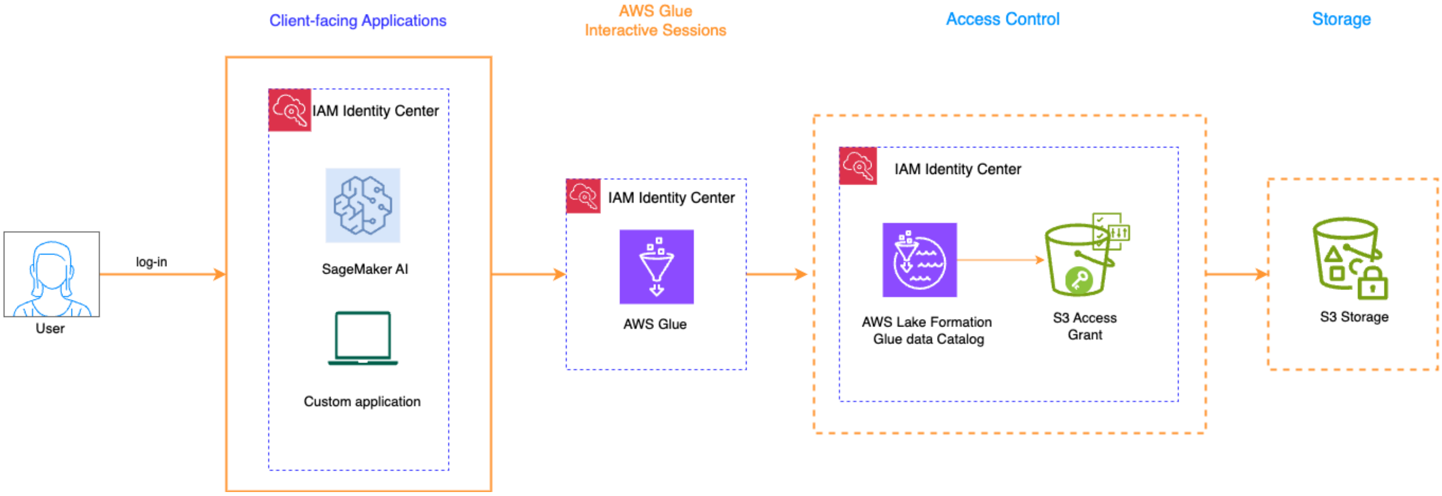
Data engineers use their corporate identities to seamlessly access and analyze data across multiple AWS accounts. Through SageMaker Studio, they launch interactive Spark sessions via AWS Glue

ETL, connecting to various data sources including Amazon S3 and the AWS Glue Data Catalog. As engineers explore datasets, Spark enforces fine-grained access controls defined in Lake Formation based on their identities, ensuring they can only view authorized data. All queries and data transformations are logged with the user's identity, creating a clear audit trail. This streamlined approach enables rapid prototyping of new analytics products while maintaining strict data governance across client environments.

Data Preparation and Feature Engineering

Data scientists from multiple research teams collaborate on complex projects using a unified data platform. They log into SageMaker Studio with their corporate credentials, immediately accessing a vast, shared data lake that spans multiple AWS accounts. As they begin feature engineering for new machine learning models, Spark sessions launched through AWS Glue ETL enforce Lake Formation's column and row-level security policies based on their propagated identities. Scientists can efficiently prepare data and engineer features using familiar tools, while compliance teams have assurance that every data interaction is automatically tracked and audited. This secure, collaborative environment accelerates research pipelines while maintaining the strict data protection standards required in regulated industries.

How it works


Client-facing Applications
AWS Glue Interactive Sessions
Access Control
Storage
User
log-in
IAM Identity Center
SageMaker AI
Custom application
IAM Identity Center
AWS Glue
IAM Identity Center
AWS Lake Formation Glue data Catalog
S3 Access Grant
S3 Storage

A user logs into client-facing applications (SageMaker AI, or custom applications) using their corporate identity through IAM Identity Center. This identity is then propagated through the entire data access pipeline.

The authenticated user launches AWS AWS Glue Interactive Sessions, which serve as the compute engine for data processing. These sessions maintain the user's identity context throughout the workflow.

AWS Lake Formation and the AWS Glue Data Catalog work together to enforce fine-grained access controls. Lake Formation applies security policies based on the user's propagated identity, while Amazon S3 Access Grant provides additional permission layers, ensuring users can only access data they're authorized to view.

Finally, the system connects to Amazon S3 Storage where the actual data resides. All access is governed by the combined security policies, maintaining data governance while enabling interactive data exploration and analysis. This architecture enables secure, identity-based data access across multiple AWS services while maintaining a seamless user experience for data scientists and engineers working with large datasets.

Integrations

AWS managed development environment

The following AWS managed client-facing applications support trusted identity propagation with AWS Glue interactive sessions:

- Sagemaker Unified Studio
- Amazon SageMaker AI

Sagemaker Unified Studio

To use trusted identity propagation with Sagemaker Unified Studio:

1. Set up Sagemaker Unified Studio project with trusted identity propagation enabled as the client-facing development environment.
2. Set up Lake Formation to enable fine-grained access control for AWS Glue tables based on the user or group in IAM Identity Center.
3. Set up Amazon S3 Access Grants to enable temporary access to the underlying data locations in Amazon S3.
4. Open Sagemaker Unified Studio JupyterLab IDE space and select AWS Glue as compute for notebook execution.

Customer managed self-hosted Notebook environment

To enable trusted identity propagation for users of custom-developed applications, see Access AWS services programmatically using trusted identity propagation in the AWS Security Blog.

Getting started with trusted identity propagation in AWS Glue ETL

This section helps you configure AWS Glue application with interactive sessions to integrate with IAM Identity Center and enable Trusted identity propagation.

Prerequisites

- An Identity Center instance in the AWS region where you want to create Trusted identity propagation enabled AWS Glue interactive sessions. An Identity Center instance can only exist in a single region for an AWS account. For more information, see Enable IAM Identity Center and provision the users and groups from your source of identities into IAM Identity Center.
- Enable Trusted identity propagation for downstream services such as Lake Formation or Amazon S3 Access Grants or Amazon Redshift cluster with which interactive workload interacts to access data.

Permissions needed to connect AWS Glue ETL with IAM Identity Center

Create an IAM role

The role that creates IAM Identity Center connection requires permissions to create and modify application configuration in AWS Glue and IAM Identity Center as in the following inline policy.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "glue:CreateGlueIdentityCenterConfiguration",
                "sso:CreateApplication",
                "sso:PutApplicationAssignmentConfiguration",
                "sso:PutApplicationAuthenticationMethod",
                "sso:PutApplicationGrant",
                "sso:PutApplicationAccessScope",
                "sso:ListInstances"
            ],
            "Resource": [
                "*"
            ]
        }
    ]
}
```

The following inline policies contain specific permissions required to view, update, and delete properties of AWS Glue integration with IAM Identity Center.

Use the following inline policy to allow an IAM role to view a AWS Glue integration with IAM Identity Center.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "glue:GetGlueIdentityCenterConfiguration"
            ],
            "Resource": [
                "*"
            ]
        }
    ]
}
```

Use the following inline policy to allow an IAM role to update AWS Glue integration with IAM Identity Center.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "glue:UpdateGlueIdentityCenterConfiguration",
                "sso:PutApplicationAccessScope",
                "sso:DeleteApplicationAccessScope"
            ],
            "Resource": [
                "*"
            ]
        }
    ]
}
```

Use the following inline policy to allow an IAM role to delete a AWS Glue integration with IAM Identity Center.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "glue:DeleteGlueIdentityCenterConfiguration",
                "sso:DeleteApplication"
            ],
            "Resource": [
                "*"
            ]
        }
    ]
}
```

Permissions description

- `glue:CreateGlueIdentityCenterConfiguration` – Grants permission to create the AWS Glue IdC configuration.
- `glue:GetGlueIdentityCenterConfiguration` – Grants permission to get an existing IdC configuration.
- `glue:DeleteGlueIdentityCenterConfiguration` – Grants permission to delete an existing AWS Glue IdC configuration.
- `glue:UpdateGlueIdentityCenterConfiguration` – Grants permission to update an existing AWS Glue IdC configuration.
- `sso:CreateApplication` – Grants permission to create a AWS Glue managed IAM Identity Center application.
- `sso:DescribeApplication` - Grants permission to describe a AWS Glue managed IAM Identity Center application.
- `sso:DeleteApplication` – Grants permission to delete a AWS Glue managed IAM Identity Center application.
- `sso:UpdateApplication` – Grants permission to update a AWS Glue managed IAM Identity Center application.

- `sso:PutApplicationGrant` – Grants permission to apply token-exchange, introspectToken, refreshToken and RevokeToken grants on IdC Application.
- `sso:PutApplicationAuthenticationMethod` – Grants permission to put authenticationMethod on AWS Glue managed IdC Application that allows AWS Glue service principal to interact with IdC Application.
- `sso:PutApplicationAccessScope` – Grants permission to add or update the list of authorized down stream service scopes on the AWS Glue managed IdC application.
- `sso:DeleteApplicationAccessScope` - Grants permission to delete downstream scopes if a scope is removed for the AWS Glue managed IdC application.
- `sso:PutApplicationAssignmentConfiguration` – Grants permission to set "User-assignment-not-required" setting on IdC Application.
- `sso:ListInstances` – Grants permission to list instances and validate the IdC InstanceArn that you specify in identity-center-configuration parameter.

Connecting AWS Glue with IAM Identity Center

When AWS Glue is connected to IAM Identity Center, it creates a singleton managed IdC application per account. The following example shows how you can connect AWS Glue with IAM Identity Center:

```
aws glue create-glue-identity-center-configuration \
--instance-arn arn:aws:sso:::instance/ssoins-123456789 \
--scopes '["s3:access_grants:read_write", "redshift:connect","lakeformation:query"]'
```

To update the scopes of the managed application (usually done to propagate to more downstream services), you can use:

```
aws glue update-glue-identity-center-configuration \
--scopes '["s3:access_grants:read_write", "redshift:connect","lakeformation:query"]'
```

Scopes parameter is optional and all scopes will be added if not provided. The supported values are `s3:access_grants:read_write`, `redshift:connect` and `lakeformation:query`.

To get the details of the configuration, you can use:

```
aws glue get-glue-identity-center-configuration
```

You can delete the connection between AWS Glue and IAM Identity Center by using the following command:

```
aws glue delete-glue-identity-center-configuration
```

Note

AWS Glue creates a service managed Identity Center Application in your account that service leverages for identity validations and identity propagation to downstream services. AWS Glue created managed Identity Center Application is shared across all trusted-identity-propagation sessions in your account.
Warning: Do not manually modify settings on the managed Identity Center Application. Any changes could affect all trusted-identity-propagation enabled AWS Glue interactive sessions in your account.

Creating a AWS Glue Interactive Session with Trusted Identity Propagation Enabled

After you connect AWS Glue with IAM Identity Center, you can use identity-enhanced role credentials to create a AWS Glue interactive session. You need not pass additional parameters when creating a 5.0 AWS Glue session. Since AWS Glue is connected with IAM identity center, if AWS Glue detects identity-enhanced-role-credentials, it will automatically propagate the identity information to downstream services which are called as part of your statements. However, the runtime role for the session needs to have the `sts:SetContext` permission as depicted below.

Runtime Role permissions to propagate identity

As AWS Glue sessions leverage Identity-enhanced credentials to propagate identity to downstream AWS services, its runtime role's trust-policy need to have addition permission `sts:SetContext` to allow identity propagation to downstream services (Amazon S3 access-grant, Lake Formation, Amazon Redshift). To learn more about how to create a runtime role, see Setting up a runtime role.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {
        "Service": "glue.amazonaws.com"
      },
      "Action": [
        "sts:AssumeRole",
        "sts:SetContext"
      ]
    }
  ]
}
```

Additionally, Runtime role would need permissions for downstream AWS services which job-run would invoke to fetch data using user identity. Please refer to the following links to configure Amazon S3 Access Grants and Lake Formation:

- Using Lake Formation with AWS Glue
- Using Amazon S3 Access Grants with AWS Glue

Considerations and limitations for AWS Glue ETL Trusted Identity Propagation integration

> ⚠ **Important**
>
> By default sessions are not private which means one IdC user can access another IdC user's session. You can use tagOnCreate to make your sessions private. For example, the session can be tagged with an owner tag and the value of it as IDC User Id and then on the policy, you can use a global condition key like `identitystore:UserId` to validate against the owner tag in the client principal/runtime role policy for all session API operations to ensure that one IdC User isn't able to access another IdC user's session.

Consider the following points when you use IAM Identity Center Trusted Identity Propagation with AWS Glue Application:

- Trusted Identity Propagation through Identity Center is supported on AWS Glue 5.0 and higher, and only with AWS Glue interactive sessions.
- AWS Glue data catalog is covered under Lake Formation identity center integration.

- Trusted Identity Propagation is limited to interactive sessions in AWS Glue, excluding other data processing entities like jobs, triggers, workflows, and ML tasks. All AWS Glue APIs, however, record user identities in AWS CloudTrail for auditing.
- AWS Glue currently supports integration with IAM Identity Center exclusively through API and CLI interfaces, not via the console.
- Once an application is enabled on AWS Glue side, make sure to create 5.0 sessions with IdC Credentials but don't create a 4.0 session with IdC credentials.
- Trusted Identity Propagation with AWS Glue is supported in the following AWS Regions:
  - af-south-1 – Africa (Cape Town)
  - ap-east-1 – Asia Pacific (Hong Kong)
  - ap-northeast-1 – Asia Pacific (Tokyo)
  - ap-northeast-2 – Asia Pacific (Seoul)
  - ap-northeast-3 – Asia Pacific (Osaka)
  - ap-south-1 – Asia Pacific (Mumbai)
  - ap-southeast-1 – Asia Pacific (Singapore)
  - ap-southeast-2 – Asia Pacific (Sydney)
  - ap-southeast-3 – Asia Pacific (Jakarta)
  - ca-central-1 – Canada (Central)
  - eu-central-1 – Europe (Frankfurt)
  - eu-north-1 – Europe (Stockholm)
  - eu-south-1 – Europe (Milan)
  - eu-west-1 – Europe (Ireland)
  - eu-west-2 – Europe (London)
  - eu-west-3 – Europe (Paris)
  - me-south-1 – Middle East (Bahrain)
  - sa-east-1 – South America (São Paulo)
  - us-east-1 – US East (N. Virginia)
  - us-east-2 – US East (Ohio)
  - us-west-1 – US West (N. California)
  - us-west-2 – US West (Oregon)

Logging and monitoring in AWS Glue

You can automate the running of your ETL (extract, transform, and load) jobs. AWS Glue provides metrics for crawlers and jobs that you can monitor. After you set up the AWS Glue Data Catalog with the required metadata, AWS Glue provides statistics about the health of your environment. You can automate the invocation of crawlers and jobs with a time-based schedule based on cron. You can also trigger jobs when an event-based trigger fires.

AWS Glue is integrated with AWS CloudTrail, a service that provides a record of actions taken by a user, role, or AWS service in AWS Glue. If you create a trail, you can enable continuous delivery of CloudTrail events to an Amazon Simple Storage Service (Amazon S3) bucket, Amazon CloudWatch Logs, and Amazon CloudWatch Events. Every event or log entry contains information about who generated the request.

Use Amazon CloudWatch Events to automate your AWS services and respond automatically to system events such as application availability issues or resource changes. Events from AWS services are delivered to CloudWatch Events in near-real time. You can write simple rules to indicate which events are of interest and what automated actions to take when an event matches a rule.

See also

- Automating AWS Glue with EventBridge
- Cross-account CloudTrail logging

An important facet of security in the cloud is logging. You must configure logging in a way that does not capture secrets and confidential material while capturing information necessary to debug and secure your cloud infastructure. Make sure to familizarize yourself with what is being logged.

Compliance validation for AWS Glue

To learn whether an AWS service is within the scope of specific compliance programs, see AWS services in Scope by Compliance Program and choose the compliance program that you are interested in. For general information, see AWS Compliance Programs.

You can download third-party audit reports using AWS Artifact. For more information, see Downloading Reports in AWS Artifact.

Your compliance responsibility when using AWS services is determined by the sensitivity of your data, your company's compliance objectives, and applicable laws and regulations. AWS provides the following resources to help with compliance:

- Security Compliance & Governance – These solution implementation guides discuss architectural considerations and provide steps for deploying security and compliance features.
- HIPAA Eligible Services Reference – Lists HIPAA eligible services. Not all AWS services are HIPAA eligible.
- AWS Compliance Resources – This collection of workbooks and guides might apply to your industry and location.
- AWS Customer Compliance Guides – Understand the shared responsibility model through the lens of compliance. The guides summarize the best practices for securing AWS services and map the guidance to security controls across multiple frameworks (including National Institute of Standards and Technology (NIST), Payment Card Industry Security Standards Council (PCI), and International Organization for Standardization (ISO)).
- Evaluating Resources with Rules in the *AWS Config Developer Guide* – The AWS Config service assesses how well your resource configurations comply with internal practices, industry guidelines, and regulations.
- AWS Security Hub – This AWS service provides a comprehensive view of your security state within AWS. Security Hub uses security controls to evaluate your AWS resources and to check your compliance against security industry standards and best practices. For a list of supported services and controls, see Security Hub controls reference.
- Amazon GuardDuty – This AWS service detects potential threats to your AWS accounts, workloads, containers, and data by monitoring your environment for suspicious and malicious activities. GuardDuty can help you address various compliance requirements, like PCI DSS, by meeting intrusion detection requirements mandated by certain compliance frameworks.
- AWS Audit Manager – This AWS service helps you continuously audit your AWS usage to simplify how you manage risk and compliance with regulations and industry standards.

Resilience in AWS Glue

The AWS global infrastructure is built around AWS Regions and Availability Zones. AWS Regions provide multiple physically separated and isolated Availability Zones, which are connected with low-latency, high-throughput, and highly redundant networking. With Availability Zones, you can design and operate applications and databases that automatically fail over between Availability

Zones without interruption. Availability Zones are more highly available, fault tolerant, and scalable than traditional single or multiple data center infrastructures.

For more information about AWS Regions and Availability Zones, see AWS Global Infrastructure.

For more information about AWS Glue job resiliency, see Error: Failover behavior between VPCs in AWS Glue.

Infrastructure security in AWS Glue

As a managed service, AWS Glue is protected by the AWS global network security procedures that are described in the Amazon Web Services: Overview of Security Processes whitepaper.

You use AWS published API calls to access AWS Glue through the network. Clients must support Transport Layer Security (TLS) 1.0 or later. We recommend TLS 1.2 or later. Clients must also support cipher suites with perfect forward secrecy (PFS) such as Ephemeral Diffie-Hellman (DHE) or Elliptic Curve Ephemeral Diffie-Hellman (ECDHE). Most modern systems such as Java 7 and later support these modes.

Additionally, requests must be signed by using an access key ID and a secret access key that is associated with an IAM principal. Or you can use the AWS Security Token Service (AWS STS) to generate temporary security credentials to sign requests.

Topics

- Configuring interface VPC endpoints (AWS PrivateLink) for AWS Glue (AWS PrivateLink)
- Configuring shared Amazon VPCs

Configuring interface VPC endpoints (AWS PrivateLink) for AWS Glue (AWS PrivateLink)

You can establish a private connection between your VPC and AWS Glue by creating an *interface VPC endpoint*. Interface endpoints are powered by AWS PrivateLink, a technology that enables you to privately access AWS Glue APIs without an internet gateway, NAT device, VPN connection, or AWS Direct Connect connection. Instances in your VPC don't need public IP addresses to communicate with AWS Glue APIs. Traffic between your VPC and AWS Glue does not leave the Amazon network.

Each interface endpoint is represented by one or more Elastic Network Interfaces in your subnets.

For more information, see Interface VPC endpoints (AWS PrivateLink) in the *Amazon VPC User Guide*.

Considerations for AWS Glue VPC endpoints

Before you set up an interface VPC endpoint for AWS Glue, ensure that you review Interface endpoint properties and limitations in the *Amazon VPC User Guide*.

AWS Glue supports making calls to all of its API actions from your VPC.

Creating an interface VPC endpoint for AWS Glue

You can create a VPC endpoint for the AWS Glue service using either the Amazon VPC console or the AWS Command Line Interface (AWS CLI). For more information, see Creating an interface endpoint in the *Amazon VPC User Guide*.

Create a VPC endpoint for AWS Glue using the following service name:

- com.amazonaws.*region*.glue

If you enable private DNS for the endpoint, you can make API requests to AWS Glue using its default DNS name for the Region, for example, `glue.us-east-1.amazonaws.com`.

For more information, see Accessing a service through an interface endpoint in the *Amazon VPC User Guide*.

Creating a VPC endpoint policy for AWS Glue

You can attach an endpoint policy to your VPC endpoint that controls access to AWS Glue. The policy specifies the following information:

- The principal that can perform actions.
- The actions that can be performed.
- The resources on which actions can be performed.

For more information, see Controlling access to services with VPC endpoints in the *Amazon VPC User Guide*.

Example: VPC endpoint policy for AWS Glue to allow job creation and update

The following is an example of an endpoint policy for AWS Glue. When attached to an endpoint, this policy grants access to the listed AWS Glue actions for all principals on all resources.

```
{
  "Statement": [
    {
      "Sid": "RestrictPassRole",
      "Effect": "Allow",
      "Action": "iam:PassRole",
      "Resource": "arn:aws:iam::123456789012:role/GlueServiceRole*",
      "Condition": {
        "StringEquals": {
          "iam:PassedToService": "glue.amazonaws.com"
        }
      }
    }
  ]
}
```

Example: VPC endpoint policy to allow read-only Data Catalog access

The following is an example of an endpoint policy for AWS Glue. When attached to an endpoint, this policy grants access to the listed AWS Glue actions for all principals on all resources.

```
{
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:GetDatabase",
        "glue:GetDatabases",
        "glue:GetTable",
        "glue:GetTables",
        "glue:GetTableVersion",
        "glue:GetTableVersions",
        "glue:GetPartition",
        "glue:GetPartitions",
        "glue:BatchGetPartition",
        "glue:SearchTables"
      ],
      "Resource": "*"
    }
  ]
}
```

Configuring shared Amazon VPCs

AWS Glue supports shared virtual private clouds (VPCs) in Amazon Virtual Private Cloud. Amazon VPC sharing allows multiple AWS accounts to create their application resources, such as Amazon EC2 instances and Amazon Relational Database Service (Amazon RDS) databases, into shared, centrally-managed Amazon VPCs. In this model, the account that owns the VPC (owner) shares one or more subnets with other accounts (participants) that belong to the same organization from AWS Organizations. After a subnet is shared, the participants can view, create, modify, and delete their application resources in the subnets that are shared with them.

In AWS Glue, to create a connection with a shared subnet, you must create a security group within your account and attach the security group to the shared subnet.

For more information, see these topics:

- Working with Shared VPCs in the *Amazon VPC User Guide*
- What Is AWS Organizations? in the *AWS Organizations User Guide*

Troubleshooting AWS Glue

If you encounter issues when working with AWS Glue, consult the topics in this section.

Topics

- Gathering AWS Glue troubleshooting information
- Troubleshooting errors Spark errors
- Crawler errors when the crawler is using Lake Formation permissions
- Troubleshooting AWS Glue for Ray errors from logs
- AWS Glue machine learning exceptions
- AWS Glue quotas

Gathering AWS Glue troubleshooting information

If you encounter errors or unexpected behavior in AWS Glue and need to contact AWS Support, you should first gather information about names, IDs, and logs that are associated with the failed action. Having this information available enables Support to help you resolve the problems you're experiencing.

Along with your *account ID*, gather the following information for each of these types of failures:

When a crawler fails, gather the following information:

- Crawler name

  Logs from crawler runs are located in CloudWatch Logs under `/aws-glue/crawlers`.

When a test connection fails, gather the following information:

- Connection name
- Connection ID
- JDBC connection string in the form `jdbc:protocol://host:port/database-name`.

  Logs from test connections are located in CloudWatch Logs under `/aws-glue/testconnection`.

When a job fails, gather the following information:

- Job name

- Job run ID in the form `jr_xxxxx`.

  Logs from job runs are located in CloudWatch Logs under `/aws-glue/jobs`.

Troubleshooting errors Spark errors

If you encounter errors in AWS Glue, use the following information to help you find the source of the problems and fix them.

> **Note**
>
> The AWS Glue GitHub repository contains additional troubleshooting guidance in AWS Glue Frequently Asked Questions.

Topics

- Error: Resource unavailable
- Error: Could not find S3 endpoint or NAT gateway for subnetId in VPC
- Error: Inbound rule in security group required
- Error: Outbound rule in security group required
- Error: Job run failed because the role passed should be given assume role permissions for the AWS Glue service
- Error: DescribeVpcEndpoints action is unauthorized. unable to validate VPC ID vpc-id
- Error: DescribeRouteTables action is unauthorized. unable to validate subnet id: Subnet-id in VPC id: vpc-id
- Error: Failed to call ec2:DescribeSubnets
- Error: Failed to call ec2:DescribeSecurityGroups
- Error: Could not find subnet for AZ
- Error: Job run exception when writing to a JDBC target
- Error: Amazon S3: The operation is not valid for the object's storage class
- Error: Amazon S3 timeout
- Error: Amazon S3 access denied
- Error: Amazon S3 access key ID does not exist

- Error: Job run fails when accessing Amazon S3 with an s3a:// URI
- Error: Amazon S3 service token expired
- Error: No private DNS for network interface found
- Error: Development endpoint provisioning failed
- Error: Notebook server CREATE_FAILED
- Error: Local notebook fails to start
- Error: Running crawler failed
- Error: Partitions were not updated
- Error: Job bookmark update failed due to version mismatch
- Error: A job is reprocessing data when job bookmarks are enabled
- Error: Failover behavior between VPCs in AWS Glue

Error: Resource unavailable

If AWS Glue returns a resource unavailable message, you can view error messages or logs to help you learn more about the issue. The following tasks describe general methods for troubleshooting.

- For any connections and development endpoints that you use, check that your cluster has not run out of elastic network interfaces.

Error: Could not find S3 endpoint or NAT gateway for subnetId in VPC

Check the subnet ID and VPC ID in the message to help you diagnose the issue.

- Check that you have an Amazon S3 VPC endpoint set up, which is required with AWS Glue. In addition, check your NAT gateway if that's part of your configuration. For more information, see Amazon VPC endpoints for Amazon S3.

Error: Inbound rule in security group required

At least one security group must open all ingress ports. To limit traffic, the source security group in your inbound rule can be restricted to the same security group.

- For any connections that you use, check your security group for an inbound rule that is self-referencing. For more information, see Setting up network access to data stores.

- When you are using a development endpoint, check your security group for an inbound rule that is self-referencing. For more information, see Setting up network access to data stores.

Error: Outbound rule in security group required

At least one security group must open all egress ports. To limit traffic, the source security group in your outbound rule can be restricted to the same security group.

- For any connections that you use, check your security group for an outbound rule that is self-referencing. For more information, see Setting up network access to data stores.
- When you are using a development endpoint, check your security group for an outbound rule that is self-referencing. For more information, see Setting up network access to data stores.

Error: Job run failed because the role passed should be given assume role permissions for the AWS Glue service

The user who defines a job must have permission for `iam:PassRole` for AWS Glue.

- When a user creates an AWS Glue job, confirm that the user's role contains a policy that contains `iam:PassRole` for AWS Glue. For more information, see Step 3: Attach a policy to users or groups that access AWS Glue.

Error: DescribeVpcEndpoints action is unauthorized. unable to validate VPC ID vpc-id

- Check the policy passed to AWS Glue for the `ec2:DescribeVpcEndpoints` permission.

Error: DescribeRouteTables action is unauthorized. unable to validate subnet id: Subnet-id in VPC id: vpc-id

- Check the policy passed to AWS Glue for the `ec2:DescribeRouteTables` permission.

Error: Failed to call ec2:DescribeSubnets

- Check the policy passed to AWS Glue for the `ec2:DescribeSubnets` permission.

Error: Failed to call ec2:DescribeSecurityGroups

- Check the policy passed to AWS Glue for the `ec2:DescribeSecurityGroups` permission.

Error: Could not find subnet for AZ

- The Availability Zone might not be available to AWS Glue. Create and use a new subnet in a different Availability Zone from the one specified in the message.

Error: Job run exception when writing to a JDBC target

When you are running a job that writes to a JDBC target, the job might encounter errors in the following scenarios:

- If your job writes to a Microsoft SQL Server table, and the table has columns defined as type `Boolean`, then the table must be predefined in the SQL Server database. When you define the job on the AWS Glue console using a SQL Server target with the option **Create tables in your data target**, don't map any source columns to a target column with data type `Boolean`. You might encounter an error when the job runs.

  You can avoid the error by doing the following:

  - Choose an existing table with the **Boolean** column.
  - Edit the `ApplyMapping` transform and map the **Boolean** column in the source to a number or string in the target.
  - Edit the `ApplyMapping` transform to remove the **Boolean** column from the source.
- If your job writes to an Oracle table, you might need to adjust the length of names of Oracle objects. In some versions of Oracle, the maximum identifier length is limited to 30 bytes or 128 bytes. This limit affects the table names and column names of Oracle target data stores.

  You can avoid the error by doing the following:

  - Name Oracle target tables within the limit for your version.

- The default column names are generated from the field names in the data. To handle the case when the column names are longer than the limit, use `ApplyMapping` or `RenameField` transforms to change the name of the column to be within the limit.

Error: Amazon S3: The operation is not valid for the object's storage class

If AWS Glue returns this error, your AWS Glue job may have been reading data from tables that have partitions across Amazon S3 storage class tiers.

- By using storage class exclusions, you can ensure that your AWS Glue jobs will work on tables that have partitions across these storage class tiers. Without exclusions, jobs that read data from these tiers fail with the following error: `AmazonS3Exception: The operation is not valid for the object's storage class`.

  For more information, see Excluding Amazon S3 storage classes.

Error: Amazon S3 timeout

If AWS Glue returns a connect timed out error, it might be because it is trying to access an Amazon S3 bucket in another AWS Region.

- An Amazon S3 VPC endpoint can only route traffic to buckets within an AWS Region. If you need to connect to buckets in other Regions, a possible workaround is to use a NAT gateway. For more information, see NAT Gateways.

Error: Amazon S3 access denied

If AWS Glue returns an access denied error to an Amazon S3 bucket or object, it might be because the IAM role provided does not have a policy with permission to your data store.

- An ETL job must have access to an Amazon S3 data store used as a source or target. A crawler must have access to an Amazon S3 data store that it crawls. For more information, see Step 2: Create an IAM role for AWS Glue.

Error: Amazon S3 access key ID does not exist

If AWS Glue returns an access key ID does not exist error when running a job, it might be because of one of the following reasons:

- An ETL job uses an IAM role to access data stores, confirm that the IAM role for your job was not deleted before the job started.
- An IAM role contains permissions to access your data stores, confirm that any attached Amazon S3 policy containing `s3:ListBucket` is correct.

Error: Job run fails when accessing Amazon S3 with an `s3a://` URI

If a job run returns an error like *Failed to parse XML document with handler class* , it might be because of a failure trying to list hundreds of files using an `s3a://` URI. Access your data store using an `s3://` URI instead. The following exception trace highlights the errors to look for:

```
1. com.amazonaws.SdkClientException: Failed to parse XML document with handler class
 com.amazonaws.services.s3.model.transform.XmlResponsesSaxParser$ListBucketHandler
2. at
 com.amazonaws.services.s3.model.transform.XmlResponsesSaxParser.parseXmlInputStream(XmlRespons
3. at
 com.amazonaws.services.s3.model.transform.XmlResponsesSaxParser.parseListBucketObjectsResponse
4. at com.amazonaws.services.s3.model.transform.Unmarshallers
$ListObjectsUnmarshaller.unmarshall(Unmarshallers.java:70)
5. at com.amazonaws.services.s3.model.transform.Unmarshallers
$ListObjectsUnmarshaller.unmarshall(Unmarshallers.java:59)
6. at
 com.amazonaws.services.s3.internal.S3XmlResponseHandler.handle(S3XmlResponseHandler.java:62)
7. at
 com.amazonaws.services.s3.internal.S3XmlResponseHandler.handle(S3XmlResponseHandler.java:31)
8. at
 com.amazonaws.http.response.AwsResponseHandlerAdapter.handle(AwsResponseHandlerAdapter.java:70
9. at com.amazonaws.http.AmazonHttpClient
$RequestExecutor.handleResponse(AmazonHttpClient.java:1554)
10. at com.amazonaws.http.AmazonHttpClient
$RequestExecutor.executeOneRequest(AmazonHttpClient.java:1272)
11. at com.amazonaws.http.AmazonHttpClient
$RequestExecutor.executeHelper(AmazonHttpClient.java:1056)
12. at com.amazonaws.http.AmazonHttpClient
$RequestExecutor.doExecute(AmazonHttpClient.java:743)
```

```
13. at com.amazonaws.http.AmazonHttpClient
$RequestExecutor.executeWithTimer(AmazonHttpClient.java:717)
14. at com.amazonaws.http.AmazonHttpClient
$RequestExecutor.execute(AmazonHttpClient.java:699)
15. at com.amazonaws.http.AmazonHttpClient$RequestExecutor.access
$500(AmazonHttpClient.java:667)
16. at com.amazonaws.http.AmazonHttpClient
$RequestExecutionBuilderImpl.execute(AmazonHttpClient.java:649)
17. at com.amazonaws.http.AmazonHttpClient.execute(AmazonHttpClient.java:513)
18. at com.amazonaws.services.s3.AmazonS3Client.invoke(AmazonS3Client.java:4325)
19. at com.amazonaws.services.s3.AmazonS3Client.invoke(AmazonS3Client.java:4272)
20. at com.amazonaws.services.s3.AmazonS3Client.invoke(AmazonS3Client.java:4266)
21. at com.amazonaws.services.s3.AmazonS3Client.listObjects(AmazonS3Client.java:834)
22. at org.apache.hadoop.fs.s3a.S3AFileSystem.getFileStatus(S3AFileSystem.java:971)
23. at
 org.apache.hadoop.fs.s3a.S3AFileSystem.deleteUnnecessaryFakeDirectories(S3AFileSystem.java:115
24. at org.apache.hadoop.fs.s3a.S3AFileSystem.finishedWrite(S3AFileSystem.java:1144)
25. at org.apache.hadoop.fs.s3a.S3AOutputStream.close(S3AOutputStream.java:142)
26. at org.apache.hadoop.fs.FSDataOutputStream
$PositionCache.close(FSDataOutputStream.java:74)
27. at org.apache.hadoop.fs.FSDataOutputStream.close(FSDataOutputStream.java:108)
28. at org.apache.parquet.hadoop.ParquetFileWriter.end(ParquetFileWriter.java:467)
29. at
 org.apache.parquet.hadoop.InternalParquetRecordWriter.close(InternalParquetRecordWriter.java:1
30. at
 org.apache.parquet.hadoop.ParquetRecordWriter.close(ParquetRecordWriter.java:112)
31. at
 org.apache.spark.sql.execution.datasources.parquet.ParquetOutputWriter.close(ParquetOutputWrit
32. at org.apache.spark.sql.execution.datasources.FileFormatWriter
$SingleDirectoryWriteTask.releaseResources(FileFormatWriter.scala:252)
33. at org.apache.spark.sql.execution.datasources.FileFormatWriter$$anonfun
$org$apache$spark$sql$execution$datasources$FileFormatWriter$$executeTask
$3.apply(FileFormatWriter.scala:191)
34. at org.apache.spark.sql.execution.datasources.FileFormatWriter$$anonfun
$org$apache$spark$sql$execution$datasources$FileFormatWriter$$executeTask
$3.apply(FileFormatWriter.scala:188)
35. at org.apache.spark.util.Utils
$.tryWithSafeFinallyAndFailureCallbacks(Utils.scala:1341)
36. at org.apache.spark.sql.execution.datasources.FileFormatWriter$.org$apache$spark
$sql$execution$datasources$FileFormatWriter$$executeTask(FileFormatWriter.scala:193)
37. at org.apache.spark.sql.execution.datasources.FileFormatWriter$$anonfun$write$1$
$anonfun$3.apply(FileFormatWriter.scala:129)
38. at org.apache.spark.sql.execution.datasources.FileFormatWriter$$anonfun$write$1$
$anonfun$3.apply(FileFormatWriter.scala:128)
```

```
39. at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:87)
40. at org.apache.spark.scheduler.Task.run(Task.scala:99)
41. at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:282)
42. at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
43. at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
44. at java.lang.Thread.run(Thread.java:748)
```

Error: Amazon S3 service token expired

When moving data to and from Amazon Redshift, temporary Amazon S3 credentials, which expire after 1 hour, are used. If you have a long running job, it might fail. For information about how to set up your long running jobs to move data to and from Amazon Redshift, see aws-glue-programming-etl-connect-redshift-home.

Error: No private DNS for network interface found

If a job fails or a development endpoint fails to provision, it might be because of a problem in the network setup.

- If you are using the Amazon provided DNS, the value of `enableDnsHostnames` must be set to true. For more information, see DNS .

Error: Development endpoint provisioning failed

If AWS Glue fails to successfully provision a development endpoint, it might be because of a problem in the network setup.

- When you define a development endpoint, the VPC, subnet, and security groups are validated to confirm that they meet certain requirements.
- If you provided the optional SSH public key, check that it is a valid SSH public key.
- Check in the VPC console that your VPC uses a valid **DHCP option set**. For more information, see DHCP option sets.
- If the cluster remains in the PROVISIONING state, contact AWS Support.

Error: Notebook server CREATE_FAILED

If AWS Glue fails to create the notebook server for a development endpoint, it might be because of one of the following problems:

- AWS Glue passes an IAM role to Amazon EC2 when it is setting up the notebook server. The IAM role must have a trust relationship to Amazon EC2.
- The IAM role must have an instance profile of the same name. When you create the role for Amazon EC2 with the IAM console, the instance profile with the same name is automatically created. Check for an error in the log regarding the instance profile name `iamInstanceProfile.name` that is not valid. For more information, see Using Instance Profiles .
- Check that your role has permission to access `aws-glue*` buckets in the policy that you pass to create the notebook server.

Error: Local notebook fails to start

If your local notebook fails to start and reports errors that a directory or folder cannot be found, it might be because of one of the following problems:

- If you are running on Microsoft Windows, make sure that the `JAVA_HOME` environment variable points to the correct Java directory. It's possible to update Java without updating this variable, and if it points to a folder that no longer exists, Jupyter notebooks fail to start.

Error: Running crawler failed

If AWS Glue fails to successfully run a crawler to catalog your data, it might be because of one of the following reasons. First check if an error is listed in the AWS Glue console crawlers list. Check if there is an exclamation icon next to the crawler name and hover over the icon to see any associated messages.

- Check the logs for the crawler run in CloudWatch Logs under `/aws-glue/crawlers`.

Error: Partitions were not updated

In case your partitions were not updated in the Data Catalog when you ran an ETL job, these log statements from the `DataSink` class in the CloudWatch logs may be helpful:

- `"Attempting to fast-forward updates to the Catalog - nameSpace:"` — Shows which database, table, and catalogId are attempted to be modified by this job. If this statement is not here, check if `enableUpdateCatalog` is set to true and properly passed as a `getSink()` parameter or in `additional_options`.

- `"Schema change policy behavior:"` — Shows which schema `updateBehavior` value you passed in.
- `"Schemas qualify (schema compare):"` — Will be true or false.
- `"Schemas qualify (case-insensitive compare):"` — Will be true or false.
- If both are false and your `updateBehavior` is not set to `UPDATE_IN_DATABASE`, then your DynamicFrame schema needs to be identical or contain a subset of the columns seen in the Data Catalog table schema.

For more information on updating partitions, see Updating the schema, and adding new partitions in the Data Catalog using AWS Glue ETL jobs.

Error: Job bookmark update failed due to version mismatch

You may be trying to parametize AWS Glue jobs to apply the same transformation/logic on different datasets in Amazon S3. You want to track processed files on the locations provided. When you run the same job on the same source bucket and write to the same/different destination concurrently (concurrency >1) the job fails with this error:

```
py4j.protocol.Py4JJavaError: An error occurred while
 callingz:com.amazonaws.services.glue.util.Job.commit.:com.amazonaws.services.gluejobexecutor.m
 Continuation update failed due to version mismatch. Expected version 2 but found
 version 3
```

Solution: set concurrency to 1 or don't run the job concurrently.

Currently AWS Glue bookmarks don't support concurrent job runs and commits will fail.

Error: A job is reprocessing data when job bookmarks are enabled

There might be cases when you have enabled AWS Glue job bookmarks, but your ETL job is reprocessing data that was already processed in an earlier run. Check for these common causes of this error:

Max Concurrency

Setting the maximum number of concurrent runs for the job greater than the default value of 1 can interfere with job bookmarks. This can occur when job bookmarks check the last modified time of objects to verify which objects need to be reprocessed. For more information, see the discussion of max concurrency in Configuring job properties for Spark jobs in AWS Glue.

Missing Job Object

Ensure that your job run script ends with the following commit:

```
job.commit()
```

When you include this object, AWS Glue records the timestamp and path of the job run. If you run the job again with the same path, AWS Glue processes only the new files. If you don't include this object and job bookmarks are enabled, the job reprocesses the already processed files along with the new files and creates redundancy in the job's target data store.

Missing Transformation Context Parameter

Transformation context is an optional parameter in the `GlueContext` class, but job bookmarks don't work if you don't include it. To resolve this error, add the transformation context parameter when you create the DynamicFrame, as shown following:

```
sample_dynF=create_dynamic_frame_from_catalog(database,
 table_name,transformation_ctx="sample_dynF")
```

Input Source

If you are using a relational database (a JDBC connection) for the input source, job bookmarks work only if the table's primary keys are in sequential order. Job bookmarks work for new rows, but not for updated rows. That is because job bookmarks look for the primary keys, which already exist. This does not apply if your input source is Amazon Simple Storage Service (Amazon S3).

Last Modified Time

For Amazon S3 input sources, job bookmarks check the last modified time of the objects, rather than the file names, to verify which objects need to be reprocessed. If your input source data has been modified since your last job run, the files are reprocessed when you run the job again.

Error: Failover behavior between VPCs in AWS Glue

The following process is used for failover for jobs in AWS Glue 4.0 and previous versions.

Summary: an AWS Glue connection is selected at the time a job run is submitted. If the job run encounters some issues, (lack of IP addresses, connectivity to source, routing problem), the job run will fail. If retries are configured, AWS Glue will retry with the same connection.

1. For each run attempt, AWS Glue will check the connections health in the order listed in the job configuration, given until it finds one it can use. In the case of an Availability Zone (AZ) failure, the connections from that AZ will fail the check and will be skipped.
2. AWS Glue validates the connection with the following:
    - checks for valid Amazon VPC id and subnet.
    - checks that a NAT gateway or Amazon VPC endpoint exists.
    - checks that the subnet has more than 0 allocated IP addresses.
    - checks that the AZ is healthy.

    AWS Glue cannot verify connectivity at the time of job run submission.
3. For jobs using Amazon VPC, all drivers and executors will be created in the same AZ with the connection selected at the time of job run submission.
4. If retries are configured, AWS Glue will retry with the same connection. This is because we cannot guarantee problems with this connection are long-running. If an AZ fails, existing job runs (depending on the stage of the job run) in that AZ can fail. A retry should detect an AZ failure and choose another AZ for the new run.

Crawler errors when the crawler is using Lake Formation permissions

Use the information below to diagnose and fix various issues while configuring the crawler using Lake Formation credentials.

Error: The S3 location: s3://examplepath is not registered

For a crawler to run using Lake Formation credentials, you need to first set up Lake Formation permissions. To resolve this error, please register the target Amazon S3 location with Lake Formation. For more information, see Registering an Amazon S3 location.

Error: User/Role is not authorized to perform: lakeformation:GetDataAccess on resource

Please add the `lakeformation:GetDataAccess` permission to the crawler role using the IAM console or AWS CLI. With this permission, Lake Formation grants the request for temporary credentials to access the data. See the policy below:

JSON

```
{
  "Version": "2012-10-17",
  "Statement": {
    "Effect": "Allow",
    "Action": [
      "lakeformation:GetDataAccess"
    ],
    "Resource": "*"
  }
}
```

Error: Insufficient Lake Formation permission(s) on (Database name: exampleDatabase, Table Name: exampleTable)

In the Lake Formation console (https://console.aws.amazon.com/lakeformation/), grant the crawler role access permissions (`Create`, `Describe`, `Alter`) on the database, which is specified as the output database. You can grant permissions on the table as well. For more information, see [Granting database permissions using the named resource method](#).

Error: Insufficient Lake Formation permission(s) on s3://examplepath

1. **Cross-account crawling**
   a. Log in to the Lake Formation console (https://console.aws.amazon.com/lakeformation/) using the account where Amazon S3 bucket is registred (account B). Grant data location permissions to the account where the crawler will be run. This will allow the crawler to read data from the target Amazon S3 location.
   b. In the account where the crawler is created (account A), grant data location permissions on the target Amazon S3 location to the IAM role used for the crawler run so that the crawler

can read the data from the destination in Lake Formation. For more information, see Granting data location permissions (external account).

2. **In-account (crawler and registered Amazon S3 location are in the same account) crawling** - Grant data location permissions to the IAM role used for the crawler run on the Amazon S3 location so that the crawler can read the data from the target in Lake Formation. For more information, see Granting data location permissions (same account).

Frequently asked questions about crawler configuration using Lake Formation credentials

1. **How do I configure a crawler to run using Lake Formation credentials using the AWS console?**

   In the AWS Glue console (https://console.aws.amazon.com/glue/), while configuring the crawler, select the option **Use Lake Formation credentials for crawling Amazon S3 data source**. For cross-account crawling, specify the AWS account ID where the target Amazon S3 location is registered with Lake Formation. For in-account crawling, the **accountId** field is optional.

2. **How do I configure a crawler to run using Lake Formation credentials using AWS CLI?**

   During `CreateCrawler` API call, add `LakeFormationConfiguration`:

   ```
   "LakeFormationConfiguration": {
       "UseLakeFormationCredentials": true,
       "AccountId": "111111111111" (AWS account ID where the target Amazon S3 location
    is registered with Lake Formation)
     }
   ```

3. **What are the supported targets for a crawler using Lake Formation credentials?**

   A crawler using Lake Formation credentials is only supported for Amazon S3 (in-account and cross-account crawling), in-account Data Catalog targets (where the underlying location is Amazon S3), and Apache Iceberg targets.

4. **Can I crawl multiple Amazon S3 buckets as part of a single crawler using Lake Formation credentials?**

   No, for crawling targets using Lake Formation credential vending, the underlying Amazon S3 locations must belong to the same bucket. For example, customers can use multiple target locations (`s3://bucket1/folder1, s3://bucket1/folder2`) if they are under the same bucket (bucket1). Specifying different buckets (s3://bucket1/folder1, s3://bucket2/folder2) is not supported.

Troubleshooting AWS Glue for Ray errors from logs

AWS Glue provides access to logs that are emitted by Ray processes during the job run. If you encounter errors or unexpected behavior in Ray jobs, first gather information from the logs to determine the cause of failure. We also provide similar logs for interactive sessions. Sessions logs are provided with the `/aws-glue/ray/sessions` prefix.

Log lines are sent to CloudWatch in real time, as your job is run. Print statements are appended to the CloudWatch logs after the run completes. Logs are retained for two weeks after a job is run.

Inspecting Ray job logs

When a job fails, gather your job name and job run ID. You can find these in the AWS Glue console. Navigate to the job page, and then navigate to the **Runs** tab. Ray job logs are stored in the following dedicated CloudWatch log groups.

- `/aws-glue/ray/jobs/script-log/` – Stores logs emitted by your main Ray script.
- `/aws-glue/ray/jobs/ray-monitor-log/` – Stores logs emitted by the Ray autoscaler process. These logs are generated for the head node and not for other worker nodes.
- `/aws-glue/ray/jobs/ray-gcs-logs/` – Stores logs emitted by the GCS (global control store) process. These logs are generated for the head node and not for other worker nodes.
- `/aws-glue/ray/jobs/ray-process-logs/` – Stores logs emitted by other Ray processes (primarily the dashboard agent) running on the head node. These logs are generated for the head node and not for other worker nodes.
- `/aws-glue/ray/jobs/ray-raylet-logs/` – Stores logs emitted by each raylet process. These logs are collected in a single stream for each worker node, including the head node.
- `/aws-glue/ray/jobs/ray-worker-out-logs/` – Stores `stdout` logs for each worker in the cluster. These logs are generated for each worker node, including the head node.
- `/aws-glue/ray/jobs/ray-worker-err-logs/` – Stores `stderr` logs for each worker in the cluster. These logs are generated for each worker node, including the head node.
- `/aws-glue/ray/jobs/ray-runtime-env-log/` – Stores logs about the Ray setup process. These logs are generated for each worker node, including the head node.

Troubleshooting Ray job errors

To understand the organization of Ray log groups, and to find the log groups that will help you troubleshoot your errors, it helps to have background information about Ray architecture.

In AWS Glue ETL, a worker corresponds to an instance. When you configure workers for an AWS Glue job, you're setting the type and quantity of instances that are dedicated to the job. Ray uses the term *worker* in different ways.

Ray uses *head node* and *worker node* to distinguish the responsibilities of an instance within a Ray cluster. A Ray worker node can host multiple *actor* processes that perform computations to achieve the result of your distributed computation. Actors that run a replica of a function are called *replicas*. Replica actors can also be called worker processes. Replicas can also run on the head node, which is known as the head because it runs additional processes to coordinate the cluster.

Each actor that contributes to your computation generates its own log stream. This provides us with some insights:

- The number of processes that emit logs might be larger than the number of workers that are allocated to the job. Often, each core on each instance has an actor.
- Ray head nodes emit cluster management and startup logs. In contrast, Ray worker nodes only emit logs for the work performed on them.

For more information about Ray architecture, see Architecture Whitepapers in the Ray documentation.

Problem area: Amazon S3 access

Check the failure message of the job run. If that doesn't provide enough information, check `/aws-glue/ray/jobs/script-log/`.

Problem area: PIP dependency management

Check `/aws-glue/ray/jobs/ray-runtime-env-log/`.

Problem area: Inspecting intermediate values in main process

Write to `stderr` or `stdout` from your main script, and retrieve logs from `/aws-glue/ray/jobs/script-log/`.

Problem area: Inspecting intermediate values in a child process

Write to `stderr` or `stdout` from your `remote` function. Then, retrieve logs from `/aws-glue/ray/jobs/ray-worker-out-logs/` or `/aws-glue/ray/jobs/ray-worker-err-logs/`. Your function might have run on any replica, so you might have to examine multiple logs to find your intended output.

Problem area: Interpreting IP addresses in error messages

In certain error situations, your job might emit an error message that contains an IP address. These IP addresses are ephemeral, and are used by the cluster to identify and communicate between nodes. Logs for a node will be published to a log stream with a unique suffix based on the IP address.

In CloudWatch, you can filter down your logs to inspect those specific to this IP address by identifying this suffix. For example, given *`FAILED_IP`* and *`JOB_RUN_ID`*, you can identify the suffix with:

```
filter @logStream like /JOB_RUN_ID/
| filter @message like /IP-/
| parse @message "IP-[*]" as ip
| filter ip like /FAILED_IP/
| fields replace(ip, ":", "_") as uIP
| stats count_distinct by uIP as logStreamSuffix
| display logStreamSuffix
```

AWS Glue machine learning exceptions

This topic describes HTTP error codes and strings for AWS Glue exceptions related to machine learning. The error codes and error strings are provided for each machine learning activity that may occur when you perform an operation. Also, you can see whether it is possible to retry the operation that resulted in the error.

CancelMLTaskRunActivity

This activity has the following exceptions:

- EntityNotFoundException (400)
  - “Cannot find MLTransform in account [accountId] with handle [transformName].”

  - “No ML Task Run found for [taskRunId]: in account [accountId] for transform [transformName].”

  OK to retry: No.

CreateMLTaskRunActivity

This activity has the following exceptions:

- InvalidInputException (400)
  - “Internal service failure due to unexpected input.”
  - “An AWS Glue Table input source should be specified in transform.”
  - “Input source column [columnName] has an invalid data type defined in the catalog.”
  - “Exactly one input record table must be provided.”
  - “Should specify database name.”
  - “Should specify table name.”
  - “Schema is not defined on the transform.”
  - “Schema should contain given primary key: [primaryKey].”
  - “Problem fetching the data catalog schema: [message].”
  - “Cannot set Max Capacity and Worker Num/Type at the same time.”
  - “Both WorkerType and NumberOfWorkers should be set.”
  - “MaxCapacity should be >= [maxCapacity].”
  - “NumberOfWorkers should be >= [maxCapacity].”
  - “Max retries should be non-negative.”
  - “Find Matches parameters have not been set.”
  - “A primary key must be specified in Find Matches parameters.”

  OK to retry: No.
- AlreadyExistsException (400)
  - “Transform with name [transformName] already exists.”

  OK to retry: No.
- IdempotentParameterMismatchException (400)

  - “Idempotent create request for transform [transformName] had mismatching parameters.”

  OK to retry: No.
- InternalServiceException (500)
  - “Dependency failure.”

  OK to retry: Yes.
- ResourceNumberLimitExceededException (400)
  - “ML Transforms count ([count]) has exceeded the limit of [limit] transforms.”

  OK to retry: Yes, once you’ve deleted a transform to make room for this new one.

DeleteMLTransformActivity

This activity has the following exceptions:

- EntityNotFoundException (400)
  - “Cannot find MLTransform in account [accountId] with handle [transformName]”

  OK to retry: No.

GetMLTaskRunActivity

This activity has the following exceptions:

- EntityNotFoundException (400)
  - “Cannot find MLTransform in account [accountId] with handle [transformName].”
  - “No ML Task Run found for [taskRunId]: in account [accountId] for transform [transformName].”

  OK to retry: No.

GetMLTaskRunsActivity

This activity has the following exceptions:

- EntityNotFoundException (400)

  - “Cannot find MLTransform in account [accountId] with handle [transformName].”
  - “No ML Task Run found for [taskRunId]: in account [accountId] for transform [transformName].”

  OK to retry: No.

GetMLTransformActivity

This activity has the following exceptions:

- EntityNotFoundException (400)
  - “Cannot find MLTransform in account [accountId] with handle [transformName].”

  OK to retry: No.

GetMLTransformsActivity

This activity has the following exceptions:

- EntityNotFoundException (400)
  - “Cannot find MLTransform in account [accountId] with handle [transformName].”

  OK to retry: No.
- InvalidInputException (400)
  - “Account ID can't be blank.”
  - “Sorting not supported for column [column].”
  - “[column] can't be blank.”
  - “Internal service failure due to unexpected input.”

  OK to retry: No.

GetSaveLocationForTransformArtifactActivity

This activity has the following exceptions:

- EntityNotFoundException (400)

  - “Cannot find MLTransform in account [accountId] with handle [transformName].”

  OK to retry: No.
- InvalidInputException (400)
  - “Unsupported artifact type [artifactType].”
  - “Internal service failure due to unexpected input.”

  OK to retry: No.

GetTaskRunArtifactActivity

This activity has the following exceptions:

- EntityNotFoundException (400)
  - “Cannot find MLTransform in account [accountId] with handle [transformName].”
  - “No ML Task Run found for [taskRunId]: in account [accountId] for transform [transformName].”

  OK to retry: No.
- InvalidInputException (400)
  - “File name ‘[fileName]’ is invalid for publish.”
  - “Cannot retrieve artifact for [taskType] task type.”
  - “Cannot retrieve artifact for [artifactType].”
  - “Internal service failure due to unexpected input.”

  OK to retry: No.

PublishMLTransformModelActivity

This activity has the following exceptions:

- EntityNotFoundException (400)
  - “Cannot find MLTransform in account [accountId] with handle [transformName].”
  - “An existing model with version - [version] cannot be found for account id - [accountId] - and transform id - [transformId].”

OK to retry: No.

- InvalidInputException (400)
  - “File name ‘[fileName]’ is invalid for publish.”
  - “Illegal leading minus sign on unsigned string [string].”
  - “Bad digit at end of [string].”
  - “String value [string] exceeds range of unsigned long.”
  - “Internal service failure due to unexpected input.”

  OK to retry: No.

PullLatestMLTransformModelActivity

This activity has the following exceptions:

- EntityNotFoundException (400)
  - “Cannot find MLTransform in account [accountId] with handle [transformName].”

  OK to retry: No.
- InvalidInputException (400)
  - “Internal service failure due to unexpected input.”

  OK to retry: No.
- ConcurrentModificationException (400)
  - “Cannot create model version to train due to racing inserts with mismatching parameters.”
  - “The ML Transform model for transform id [transformId] is stale or being updated by another process; Please retry.”

  OK to retry: Yes.

PutJobMetadataForMLTransformActivity

This activity has the following exceptions:

- EntityNotFoundException (400)
  - “Cannot find MLTransform in account [accountId] with handle [transformName].”

  - “No ML Task Run found for [taskRunId]: in account [accountId] for transform [transformName].”

  OK to retry: No.
- InvalidInputException (400)
  - “Internal service failure due to unexpected input.”
  - “Unknown job metadata type [jobType].”
  - “Must provide a task run ID to update.”

  OK to retry: No.

StartExportLabelsTaskRunActivity

This activity has the following exceptions:

- EntityNotFoundException (400)
  - “Cannot find MLTransform in account [accountId] with handle [transformName].”
  - “No labelset exists for transformId [transformId] in account id [accountId].”

  OK to retry: No.
- InvalidInputException (400)
  - “[message].”
  - “S3 path provided is not in the same region as transform. Expecting region - [region], but got - [region].”

  OK to retry: No.

StartImportLabelsTaskRunActivity

This activity has the following exceptions:

- EntityNotFoundException (400)
  - “Cannot find MLTransform in account [accountId] with handle [transformName].”

  OK to retry: No.
- InvalidInputException (400)

  - “[message].”
  - “Invalid label file path.”
  - “Cannot access the label file at [labelPath]. [message].”
  - “Cannot use IAM role provided in the transform. Role: [role].”
  - “Invalid label file of size 0.”
  - “S3 path provided is not in the same region as transform. Expecting region - [region], but got - [region].”

  OK to retry: No.
- ResourceNumberLimitExceededException (400)
  - “Label file has exceeded the limit of [limit] MB.”

  OK to retry: No. Consider breaking your label file into several smaller files.

StartMLEvaluationTaskRunActivity

This activity has the following exceptions:

- EntityNotFoundException (400)
  - “Cannot find MLTransform in account [accountId] with handle [transformName].”

  OK to retry: No.
- InvalidInputException (400)
  - “Exactly one input record table must be provided.”
  - “Should specify database name.”
  - “Should specify table name.”
  - “Find Matches parameters have not been set.”
  - “A primary key must be specified in Find Matches parameters.”

  OK to retry: No.
- MLTransformNotReadyException (400)
  - “This operation can only be applied to a transform that is in a READY state.”

  OK to retry: No.
- InternalServiceException (500)

  - “Dependency failure.”

  OK to retry: Yes.

- ConcurrentRunsExceededException (400)
  - “ML Task Runs count [count] has exceeded the transform limit of [limit] task runs.”
  - “ML Task Runs count [count] has exceeded the limit of [limit] task runs.”

  OK to retry: Yes, after waiting for task runs to finish.

StartMLLabelingSetGenerationTaskRunActivity

This activity has the following exceptions:

- EntityNotFoundException (400)
  - “Cannot find MLTransform in account [accountId] with handle [transformName].”

  OK to retry: No.

- InvalidInputException (400)
  - “Exactly one input record table must be provided.”
  - “Should specify database name.”
  - “Should specify table name.”
  - “Find Matches parameters have not been set.”
  - “A primary key must be specified in Find Matches parameters.”

  OK to retry: No.

- InternalServiceException (500)
  - “Dependency failure.”

  OK to retry: Yes.

- ConcurrentRunsExceededException (400)
  - “ML Task Runs count [count] has exceeded the transform limit of [limit] task runs.”

  OK to retry: Yes, after task runs have completed.

UpdateMLTransformActivity

This activity has the following exceptions:

- EntityNotFoundException (400)
  - “Cannot find MLTransform in account [accountId] with handle [transformName].”

  OK to retry: No.
- InvalidInputException (400)
  - “Another transform with name [transformName] already exists.”
  - “[message].”
  - “Transform name cannot be blank.”
  - “Cannot set Max Capacity and Worker Num/Type at the same time.”
  - “Both WorkerType and NumberOfWorkers should be set.”
  - “MaxCapacity should be >= [minMaxCapacity].”
  - “NumberOfWorkers should be >= [minNumWorkers].”
  - “Max retries should be non-negative.”
  - “Internal service failure due to unexpected input.”
  - “Find Matches parameters have not been set.”
  - “A primary key must be specified in Find Matches parameters.”

  OK to retry: No.
- AlreadyExistsException (400)
  - “Transform with name [transformName] already exists.”

  OK to retry: No.
- IdempotentParameterMismatchException (400)
  - “Idempotent create request for transform [transformName] had mismatching parameters.”

  OK to retry: No.

AWS Glue quotas

You can contact AWS Support to request a quota increase for the service quotas listed in the *AWS General Reference*. Unless otherwise noted, each quota is Region-specific. For more information, see AWS Glue Endpoints and Quotas.

Improving AWS Glue performance

Baseline strategy for performance tuning

In order to improve AWS Glue performance, you may consider updating certain performance related AWS Glue parameters. When preparing to tune parameters, use the following best practices:

- Determine your performance goals before beginning to identify problems.
- Use metrics to identify problems before attempting to change tuning parameters.

For the most consistent results when tuning a job, develop a baseline strategy for your tuning work.

Generally, performance tuning is performed in the following workflow:

1. Determine performance goals.
2. Measure metrics.
3. Identify bottlenecks.
4. Reduce the impact of the bottlenecks.
5. Repeat steps 2-4 until you achieve the intended target.

Tuning strategies for your job type

Spark jobs–follow the guidance in Best practices for performance tuning AWS Glue for Apache Spark jobs on AWS Prescriptive Guidance.

Other jobs–you can tune AWS Glue for Ray and AWS Glue Python shell jobs by adapting strategies available in other runtime environments.

Improving performance for AWS Glue for Apache Spark jobs

In order to improve AWS Glue for Spark performance, you may consider updating certain performance related AWS Glue and Spark parameters.

For more information about specific strategies for identifying bottlenecks through metrics and reducing their impact, see Best practices for performance tuning AWS Glue for Apache Spark jobs on AWS Prescriptive Guidance. This guide introduces you to key topics applicable to Apache Spark in all runtime environments, such as Spark architecture and Resilient Distributed Datasets. Using those topics, the guide guides you to implement specific performance tuning strategies, such as optimizing shuffles and parallelizing tasks.

You can identify bottlenecks by configuring AWS Glue to show the Spark UI. For more information, see the section called “Monitoring with the Spark UI”.

Additionally, AWS Glue provides performance features that may be applicable to the specific type of data store your job connects to. Reference information about performance parameters for data stores can be found in the section called “Connection parameters”.

Optimizing reads with pushdown in AWS Glue ETL

Pushdown is an optimization technique that pushes logic about retrieving data closer to the source of your data. The source could be a database or a file system such as Amazon S3. When executing certain operations directly on the source, you can save time and processing power by not bringing all the data over the network to the Spark engine managed by AWS Glue.

Another way of saying this is that pushdown reduces data scan. For more information about the process of identifying when this technique is appropriate, consult Reduce the amount of data scan in the *Best practices for performance tuning AWS Glue for Apache Spark jobs* guide on AWS Prescriptive Guidance.

Predicate pushdown on files stored on Amazon S3

When working with files on Amazon S3 that have been organized by prefix, you can filter your target Amazon S3 paths by defining a pushdown predicate. Rather than reading the complete dataset and applying filters within a `DynamicFrame`, you can directly apply the filter to the partition metadata stored in the AWS Glue Data Catalog. This approach allows you to selectively list and read only the necessary data. For more information about this process, including writing to a bucket by partitions, see the section called “Managing partitions”.

You achieve predicate pushdown in Amazon S3 by using the `push_down_predicate` parameter. Consider a bucket in Amazon S3 you've partitioned by year, month and day. If you want to retrieve customer data for June of 2022, you can instruct AWS Glue to read only relevant Amazon S3 paths. The `push_down_predicate` in this case is `year='2022' and month='06'`. Putting it all together, the read operation can be achieved as below:

Python

```
customer_records = glueContext.create_dynamic_frame.from_catalog(
    database = "customer_db",
    table_name = "customer_tbl",
    push_down_predicate = "year='2022' and month='06'"
)
```

Scala

```
val customer_records = glueContext.getCatalogSource(
database="customer_db",
tableName="customer_tbl",
pushDownPredicate="year='2022' and month='06'"
).getDynamicFrame()
```

In the previous scenario, `push_down_predicate` retrieves a list of all partitions from the AWS Glue Data Catalog and filters them before reading the underlying Amazon S3 files. Though this helps in most cases, when working with datasets that have millions of partitions, the process of listing partitions can be time consuming. To address this issue, server-side pruning of partitions can be used to improve performance. This is done by building a **Partition index** for your data in the AWS Glue Data Catalog. For more information about partition indices, see the section called "Creating partition indexes". You can then use the `catalogPartitionPredicate` option to reference the index. For an example retrieving partitions with `catalogPartitionPredicate`, see the section called "Catalog partition predicates".

Pushdown when working with JDBC sources

The AWS Glue JDBC reader used in the `GlueContext` supports pushdown on supported databases by providing custom SQL queries that can run directly on the source. This can be achieved by setting the `sampleQuery` parameter. Your sample query can specify which columns to select as well as provide a pushdown predicate to limit the data transferred to the Spark engine.

By default, sample queries operate on a single node, which can result in job failures when dealing with large data volumes. To use this feature to query data at scale, you should configure query partitioning by setting `enablePartitioningForSampleQuery` to true, which will distribute the query to multiple nodes across a key of your choice. Query partitioning also requires a few other necessary configuration parameters. For more information about query partitioning, see the section called "Reading from JDBC in parallel".

When setting `enablePartitioningForSampleQuery`, AWS Glue will combine your pushdown predicate with a partitioning predicate when querying your database. Your `sampleQuery` must end with an `AND` for AWS Glue to append partitioning conditions. (If you do not provide a pushdown predicate, `sampleQuery` must end with an `WHERE`). See an example below, where we push down a predicate to only retrieve rows whose `id` is greater than 1000. This `sampleQuery` will only return the name and location columns for rows where `id` is greater than the specified value:

Python

```
sample_query = "select name, location from customer_tbl WHERE id>=1000 AND"
customer_records = glueContext.create_dynamic_frame.from_catalog(
    database="customer_db",
    table_name="customer_tbl",
    sample_query = "select name, location from customer_tbl WHERE id>=1000 AND",

    additional_options = {
                           "hashpartitions": 36 ,
                           "hashfield":"id",
                           "enablePartitioningForSampleQuery":True,
                           "sampleQuery":sample_query
                          }
)
```

Scala

```
val additionalOptions = Map(
        "hashpartitions" -> "36",
        "hashfield" -> "id",
        "enablePartitioningForSampleQuery" -> "true",
        "sampleQuery" -> "select name, location from customer_tbl WHERE id >= 1000
 AND"
        )

    val customer_records = glueContext.getCatalogSource(
        database="customer_db",
        tableName="customer_tbl").getDynamicFrame()
```

> **Note**
>
> If `customer_tbl` has a different name in your Data Catalog and underlying datastore, you must provide the underlying table name in sample_query, since the query is passed to the underlying datastore.

You can also query against JDBC tables without integrating with the AWS Glue Data Catalog. Instead of providing username and password as parameters to the method, you can reuse credentials from a preexisting connection by providing `useConnectionProperties` and `connectionName`. In this example, we retrieve credentials from a connection called `my_postgre_connection`.

Python

```
connection_options_dict = {
    "useConnectionProperties": True,
    "connectionName": "my_postgre_connection",
    "dbtable":"customer_tbl",
    "sampleQuery":"select name, location from customer_tbl WHERE id>=1000 AND",
    "enablePartitioningForSampleQuery":True,
    "hashfield":"id",
    "hashpartitions":36
    }

customer_records = glueContext.create_dynamic_frame.from_options(
    connection_type="postgresql",
    connection_options=connection_options_dict
    )
```

Scala

```
val connectionOptionsJson = """
      {
        "useConnectionProperties": true,
        "connectionName": "my_postgre_connection",
        "dbtable": "customer_tbl",
        "sampleQuery": "select name, location from customer_tbl WHERE id>=1000 AND",
        "enablePartitioningForSampleQuery" : true,
        "hashfield" : "id",
        "hashpartitions" : 36
      }
    """

    val connectionOptions = new JsonOptions(connectionOptionsJson)

    val dyf = glueContext.getSource("postgresql",
  connectionOptions).getDynamicFrame()
```

Notes and limitations for pushdown in AWS Glue

Pushdown, as a concept, is applicable when reading from non-streaming sources. AWS Glue supports a variety of sources - the ability to pushdown depends on the source and connector.

- When connecting to Snowflake, you can use the `query` option. Similar functionality exists in the Redshift connector in AWS Glue 4.0 and later versions. For more information about reading from Snowflake with `query`, see the section called "Read from Snowflake".
- The DynamoDB ETL reader does not support filters or pushdown predicates. MongoDB and DocumentDB also do not support this sort of functionality.
- When reading from data stored in Amazon S3 in open table formats, the partitioning method for files in Amazon S3 is no longer sufficient. To read and write from partitions using open table formats, consult documentation for the format.
- DynamicFrame methods do not perform Amazon S3 projection pushdown. All columns will be read from files that pass the predicate filter.
- When working with `custom.jdbc` connectors in AWS Glue, the ability to pushdown depends on the source and connector. Please review the appropriate connector documentation to confirm if and how it supports pushdown in AWS Glue.

Using auto scaling for AWS Glue

Auto Scaling is available for your AWS Glue ETL, interactive sessions, and streaming jobs with AWS Glue version 3.0 or later.

With Auto Scaling enabled, you will get the following benefits:

- AWS Glue automatically adds and removes workers from the cluster depending on the parallelism at each stage or microbatch of the job run.

- It reduces the need for you to experiment and decide on the number of workers to assign for your AWS Glue ETL jobs.
- With the given maximum number of workers, AWS Glue will choose the right size resources for the workload.
- You can see how the size of the cluster changes during the job run by looking at CloudWatch metrics on the job run details page in AWS Glue Studio.

Auto Scaling for AWS Glue ETL and streaming jobs enables on-demand scaling-out and scaling-in of the computing resources of your AWS Glue jobs. On-demand scale-up helps you to only allocate the required computing resources initially on job run startup, and also to provision the required resources as per demand during the job.

Auto Scaling also supports dynamic scale-in of the AWS Glue job resources over the course of a job. Over a job run, when more executors are requested by your Spark application, more workers will be added to the cluster. When the executor has been idle without active computation tasks, the executor and the corresponding worker will be removed.

Common scenarios where Auto Scaling helps with cost and utilization for your Spark applications include:

- a Spark driver listing a large number of files in Amazon S3 or performing a load while executors are inactive
- Spark stages running with only a few executors due to over-provisioning
- data skews or uneven computation demand across Spark stages

Requirements

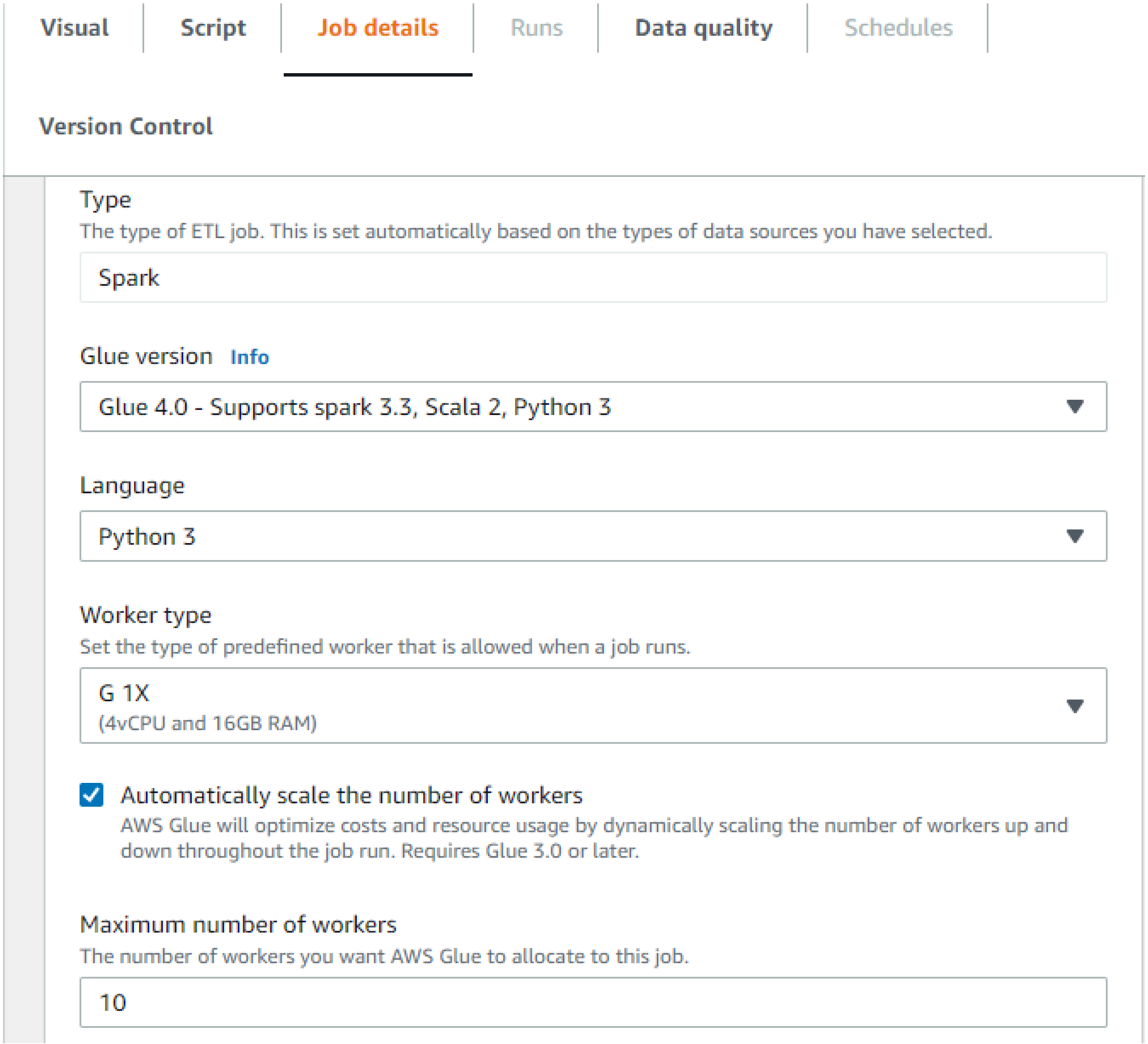
Auto Scaling is only available for AWS Glue version 3.0 or later. To use Auto Scaling, you can follow the migration guide to migrate your existing jobs to AWS Glue version 3.0 or later or create new jobs with AWS Glue version 3.0 or later.

Auto Scaling is available for AWS Glue jobs with the `G.1X`, `G.2X`, `G.4X`, `G.8X`, `G.12X`, `G.16X`, `R.1X`, `R.2X`, `R.4X`, `R.8X`, or `G.025X` (only for Streaming jobs) worker types. Standard DPUs are not supported for Auto Scaling.

Enabling Auto Scaling in AWS Glue Studio

On the **Job details** tab in AWS Glue Studio, choose the type as **Spark** or **Spark Streaming**, and **Glue version** as `Glue 3.0` or later. Then, a check box will appear below **Worker type**.

- Select the **Automatically scale the number of workers** option.
- Set the **Maximum number of workers** to define the maximum number of workers that can be vended to the job run.



Type
The type of ETL job. This is set automatically based on the types of data sources you have selected.
Spark

Glue version Info
Glue 4.0 - Supports spark 3.3, Scala 2, Python 3

Language
Python 3

Worker type
Set the type of predefined worker that is allowed when a job runs.
G 1X
(4vCPU and 16GB RAM)

☑ Automatically scale the number of workers
AWS Glue will optimize costs and resource usage by dynamically scaling the number of workers up and down throughout the job run. Requires Glue 3.0 or later.

Maximum number of workers
The number of workers you want AWS Glue to allocate to this job.
10

Enabling Auto Scaling with the AWS CLI or SDK

To enable Auto Scaling from the AWS CLI for your job run, run `start-job-run` with the following configuration:

```
{
    "JobName": "<your job name>",
    "Arguments": {
        "--enable-auto-scaling": "true"
    },
    "WorkerType": "G.2X", // G.1X, G.2X, G.4X, G.8X, G.12X, G.16X, R.1X, R.2X, R.4X,
 and R.8X are supported for Auto Scaling Jobs
    "NumberOfWorkers": 20, // represents Maximum number of workers
    ...other job run configurations...
}
```

Once at ETL job run is finished, you can also call `get-job-run` to check the actual resource usage of the job run in DPU-seconds. Note: the new field **DPUSeconds** will only show up for your batch jobs on AWS Glue 4.0 or later enabled with Auto Scaling. This field is not supported for streaming jobs.

```
$ aws glue get-job-run --job-name your-job-name --run-id jr_xx --endpoint https://
glue.us-east-1.amazonaws.com --region us-east-1
{
    "JobRun": {
        ...
        "GlueVersion": "3.0",
        "DPUSeconds": 386.0
    }
}
```

You can also configure job runs with Auto Scaling using the AWS Glue SDK with the same configuration.

Enabling Auto Scaling with Interactive sessions

To enable Auto Scaling when building AWS Glue jobs with interactive sessions, see Configuring AWS Glue interactive sessions.

Tips and considerations

Tips and the considerations for fine-tuning AWS Glue Auto Scaling:

- In case you do not have any idea on the initial value of the maximum number of the workers, you can start from the rough calculation explained in Estimate AWS Glue DPU . You should not configure extremely large value in the maximum number of workers for very low volume data.
- AWS Glue Auto Scaling configures `spark.sql.shuffle.partitions` and `spark.default.parallelism` based on the maximum number of DPU (calculated with the maximum number of workers and the worker type) configured on the job. In case you prefer the fixed value on those configurations, you can overwrite these parameters with the following job parameters:
  - **Key**: `--conf`
  - **Value**: `spark.sql.shuffle.partitions=200 --conf spark.default.parallelism=200`
- For streaming jobs, by default, AWS Glue does not auto scale within micro batches and it requires several micro batches to initiate auto scaling. In case you want to enable auto scaling within micro batches, provide `--auto-scale-within-microbatch`. For more information, see Job parameter reference .

Monitoring Auto Scaling with Amazon CloudWatch metrics

The CloudWatch executor metrics are available for your AWS Glue 3.0 or later jobs if you enable Auto Scaling. The metrics can be used to monitor the demand and optimized usage of executors in their Spark applications enabled with Auto Scaling. For more information, see Monitoring AWS Glue using Amazon CloudWatch metrics.

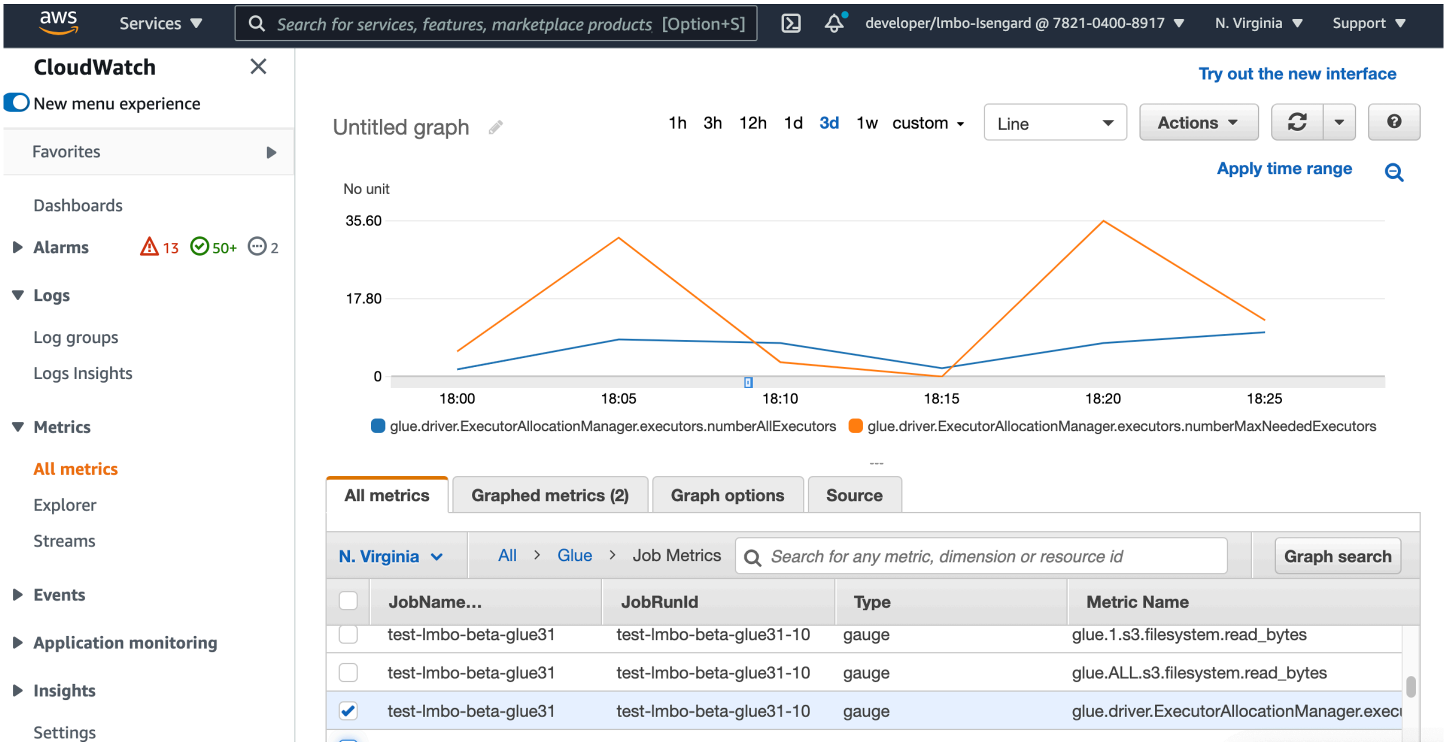
You can also utilize AWS Glue observability metrics to get insights about resource utilization. For example, by monitoring `glue.driver.workerUtilization`, you can monitor how much resource was actually used with and without auto scaling. For another example, by monitoring `glue.driver.skewness.job` and `glue.driver.skewness.stage`, you can see how the data is skewed. Those insights will help you decide to enable auto scaling and fine tune the configurations. For more information, see Monitoring with Monitoring with AWS Glue Observability metrics.

- glue.driver.ExecutorAllocationManager.executors.numberAllExecutors

- glue.driver.ExecutorAllocationManager.executors.numberMaxNeededExecutors

For more details on these metrics, see [Monitoring for DPU capacity planning](#).

> **Note**
>
> CloudWatch executor metrics are not available for interactive sessions.


aws
Services ▼
Search for services, features, marketplace products [Option+S]
developer/lmbo-Isengard @ 7821-0400-8917 ▼
N. Virginia ▼
Support ▼
CloudWatch
New menu experience
Favorites
Dashboards
Alarms 13 50+ 2
Logs
Log groups
Logs Insights
Metrics
All metrics
Explorer
Streams
Events
Application monitoring
Insights
Settings
Try out the new interface
Untitled graph
1h 3h 12h 1d 3d 1w custom
Line
Actions
Apply time range
No unit
35.60
17.80
0
18:00
18:05
18:10
18:15
18:20
18:25
glue.driver.ExecutorAllocationManager.executors.numberAllExecutors
glue.driver.ExecutorAllocationManager.executors.numberMaxNeededExecutors
All metrics
Graphed metrics (2)
Graph options
Source
N. Virginia
All > Glue > Job Metrics
Search for any metric, dimension or resource id
Graph search

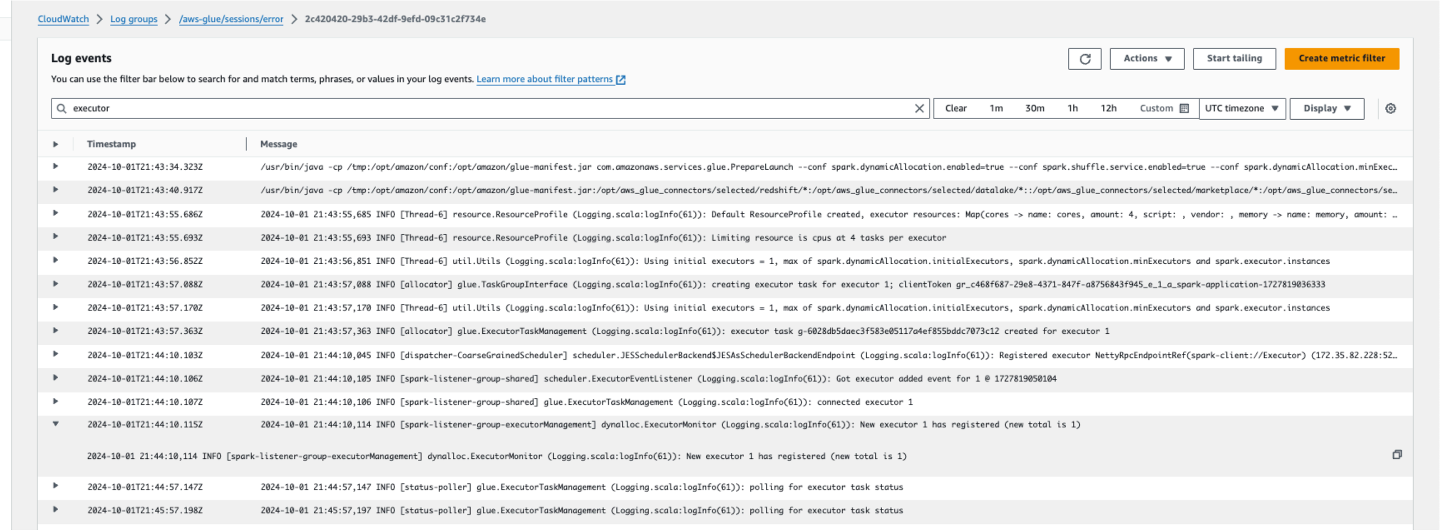
☐	JobName...	JobRunId	Type	Metric Name
☐	test-lmbo-beta-glue31	test-lmbo-beta-glue31-10	gauge	glue.1.s3.filesystem.read_bytes
☐	test-lmbo-beta-glue31	test-lmbo-beta-glue31-10	gauge	glue.ALL.s3.filesystem.read_bytes
☑	test-lmbo-beta-glue31	test-lmbo-beta-glue31-10	gauge	glue.driver.ExecutorAllocationManager.execu

Monitoring Auto Scaling with Amazon CloudWatch Logs

If you are using interactive sessions, you can monitor the number of executors by enabling continuous Amazon CloudWatch Logs and searching “executor” in the logs, or by using the Spark UI. To do this, use the `%%configure` magic to enable continuous logging along with `enable auto scaling`.

```
%%configure{
    "--enable-continuous-cloudwatch-log": "true",
    "--enable-auto-scaling": "true"
}
```

In the Amazon CloudWatch Logsevents, search "executor"in the logs:



CloudWatch > Log groups > /aws-glue/sessions/error > 2c420420-29b3-42df-9efd-09c31c2f734e

Log events

You can use the filter bar below to search for and match terms, phrases, or values in your log events. Learn more about filter patterns

executor

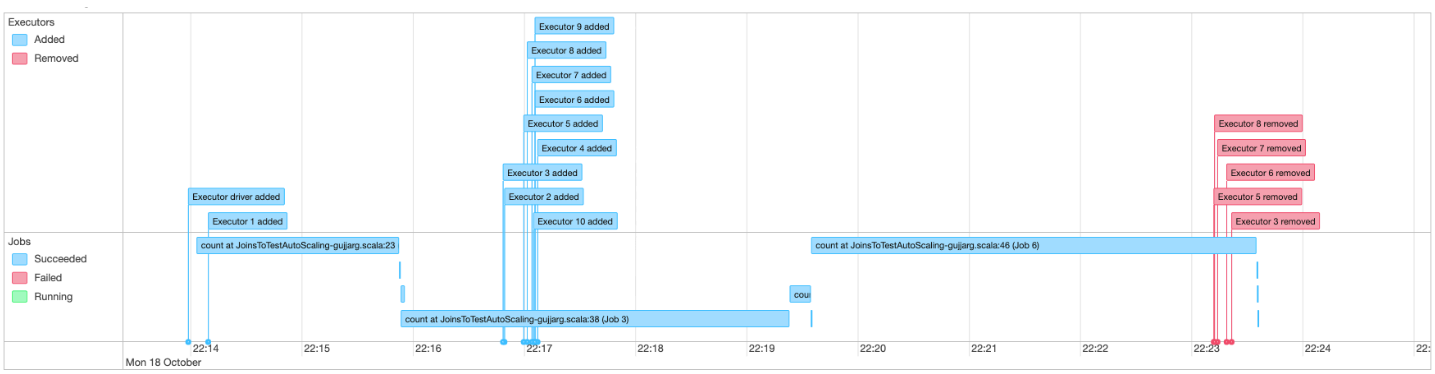
```
Timestamp                  Message
2024-10-01T21:43:34.323Z   /usr/bin/java -cp /tmp:/opt/amazon/conf:/opt/amazon/glue-manifest.jar com.amazonaws.services.glue.PrepareLaunch --conf spark.dynamicAllocation.enabled=true --conf spark.shuffle.service.enabled=true --conf spark.dynamicAllocation.minExec…
2024-10-01T21:43:40.917Z   /usr/bin/java -cp /tmp:/opt/amazon/conf:/opt/amazon/glue-manifest.jar:/opt/aws_glue_connectors/selected/redshift/*:/opt/aws_glue_connectors/selected/datalake/*::/opt/aws_glue_connectors/selected/marketplace/*:/opt/aws_glue_connectors/se…
2024-10-01T21:43:55.686Z   2024-10-01 21:43:55,685 INFO [Thread-6] resource.ResourceProfile (Logging.scala:logInfo(61)): Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 4, script: , vendor: , memory -> name: memory, amount: …
2024-10-01T21:43:55.693Z   2024-10-01 21:43:55,693 INFO [Thread-6] resource.ResourceProfile (Logging.scala:logInfo(61)): Limiting resource is cpus at 4 tasks per executor
2024-10-01T21:43:56.852Z   2024-10-01 21:43:56,851 INFO [Thread-6] util.Utils (Logging.scala:logInfo(61)): Using initial executors = 1, max of spark.dynamicAllocation.initialExecutors, spark.dynamicAllocation.minExecutors and spark.executor.instances
2024-10-01T21:43:57.088Z   2024-10-01 21:43:57,088 INFO [allocator] glue.TaskGroupInterface (Logging.scala:logInfo(61)): creating executor task for executor 1; clientToken gr_c468f687-29e8-4371-847f-a8756843f945_e_1_a_spark-application-1727819036333
2024-10-01T21:43:57.170Z   2024-10-01 21:43:57,170 INFO [Thread-6] util.Utils (Logging.scala:logInfo(61)): Using initial executors = 1, max of spark.dynamicAllocation.initialExecutors, spark.dynamicAllocation.minExecutors and spark.executor.instances
2024-10-01T21:43:57.363Z   2024-10-01 21:43:57,363 INFO [allocator] glue.ExecutorTaskManagement (Logging.scala:logInfo(61)): executor task g-6028db5daec3f583e05117a4ef855bddc7073c12 created for executor 1
2024-10-01T21:44:10.103Z   2024-10-01 21:44:10,045 INFO [dispatcher-CoarseGrainedScheduler] scheduler.JESSchedulerBackend$JESAsSchedulerBackendEndpoint (Logging.scala:logInfo(61)): Registered executor NettyRpcEndpointRef(spark-client://Executor) (172.35.82.228:52…
2024-10-01T21:44:10.106Z   2024-10-01 21:44:10,105 INFO [spark-listener-group-shared] scheduler.ExecutorEventListener (Logging.scala:logInfo(61)): Got executor added event for 1 @ 1727819050104
2024-10-01T21:44:10.107Z   2024-10-01 21:44:10,106 INFO [spark-listener-group-shared] glue.ExecutorTaskManagement (Logging.scala:logInfo(61)): connected executor 1
2024-10-01T21:44:10.115Z   2024-10-01 21:44:10,114 INFO [spark-listener-group-executorManagement] dynalloc.ExecutorMonitor (Logging.scala:logInfo(61)): New executor 1 has registered (new total is 1)

2024-10-01 21:44:10,114 INFO [spark-listener-group-executorManagement] dynalloc.ExecutorMonitor (Logging.scala:logInfo(61)): New executor 1 has registered (new total is 1)

2024-10-01T21:44:57.147Z   2024-10-01 21:44:57,147 INFO [status-poller] glue.ExecutorTaskManagement (Logging.scala:logInfo(61)): polling for executor task status
2024-10-01T21:45:57.198Z   2024-10-01 21:45:57,197 INFO [status-poller] glue.ExecutorTaskManagement (Logging.scala:logInfo(61)): polling for executor task status
```

Monitoring Auto Scaling with Spark UI

With Auto Scaling enabled, you can also monitor executors being added and removed with dynamic scale-up and scale-down based on the demand in your AWS Glue jobs using the Glue Spark UI. For more information, see Enabling the Apache Spark web UI for AWS Glue jobs.

When you use interactive sessions from Jupyter notebook, you can run following magic to enable auto scaling along with Spark UI:

```
%%configure{
    "--enable-auto-scaling": "true",
    "--enable-continuous-cloudwatch-log": "true"
}
```



Monitoring Auto Scaling job run DPU usage

You may use the AWS Glue Studio Job run view to check the DPU usage of your Auto Scaling jobs.

1. Choose **Monitoring** from the AWS Glue Studio navigation pane. The Monitoring page appears.
2. Scroll down to the Job runs chart.
3. Navigate to the job run you are interested and scroll to the DPU hours column to check the usage for the specific job run.

Limitations

AWS Glue streaming Auto Scaling currently doesn't support a streaming DataFrame join with a static DataFrame created outside of `ForEachBatch`. A static DataFrame created inside the `ForEachBatch` will work as expected.

Workload partitioning with bounded execution

Errors in Spark applications commonly arise from inefficient Spark scripts, distributed in-memory execution of large-scale transformations, and dataset abnormalities. There are many reasons that may cause driver or executor out of memory issues, for example a data skew, listing too many objects, or large data shuffles. These issues often appear when you are processing huge amounts of backlog data with Spark.

AWS Glue allows you to solve OOM issues and make your ETL processing easier with workload partitioning. With workload partitioning enabled, each ETL job run only picks unprocessed data, with an upper bound on the dataset size or the number of files to be processed with this job run. Future job runs will process the remaining data. For example, if there are 1000 files need to be processed, you can set the number of files to be 500 and separate them into two job runs.

Workload partitioning is supported only for Amazon S3 data sources.

Enabling workload partitioning

You can enable bounded execution by manually setting the options in your script or by adding catalog table properties.

To enable workload partitioning with bounded execution in your script:

1. To avoid reprocessing data, enable job bookmarks in the new job or existing job. For more information, see Tracking Processed Data Using Job Bookmarks.
2. Modify your script and set the bounded limit in the additional options in the AWS Glue `getSource` API. You should also set the transformation context for the job bookmark to store the `state` element. For example:

   Python

   ```
   glueContext.create_dynamic_frame.from_catalog(
       database = "database",
       table_name = "table_name",
       redshift_tmp_dir = "",
       transformation_ctx = "datasource0",
       additional_options = {
           "boundedFiles" : "500", # need to be string
         # "boundedSize" : "1000000000" unit is byte
       }
   )
   ```

   Scala

   ```
   val datasource0 = glueContext.getCatalogSource(
       database = "database", tableName = "table_name", redshiftTmpDir = "",
       transformationContext = "datasource0",
       additionalOptions = JsonOptions(
           Map("boundedFiles" -> "500") // need to be string
         //"boundedSize" -> "1000000000" unit is byte
       )
   ).getDynamicFrame()
   ```

   ```
   val connectionOptions = JsonOptions(
       Map("paths" -> List(baseLocation), "boundedFiles" -> "30")
   )
   val source = glueContext.getSource("s3", connectionOptions, "datasource0", "")
   ```

To enable workload partitioning with bounded execution in your Data Catalog table:

1. Set the key-value pairs in the `parameters` field of your table structure in the Data Catalog. For more information, see [Viewing and Editing Table Details].
2. Set the upper limit for the dataset size or the number of files processed:
    - Set `boundedSize` to the target size of the dataset in bytes. The job run will stop after reaching the target size from the table.
    - Set `boundedFiles` to the target number of files. The job run will stop after processing the target number of files.

> **Note**
>
> You should only set one of `boundedSize` or `boundedFiles`, as only a single boundary is supported.

Setting up an AWS Glue trigger to automatically run the job

Once you have enabled bounded execution, you can set up an AWS Glue trigger to automatically run the job and incrementally load the data in sequential runs. Go to the AWS Glue Console and create a trigger, setup the schedule time, and attach to your job. Then it will automatically trigger the next job run and process the new batch of data.

You can also use AWS Glue workflows to orchestrate multiple jobs to process data from different partitions in parallel. For more information, see AWS Glue Triggers and AWS Glue Workflows.

For more information on use cases and options, please refer to the blog Optimizing Spark applications with workload partitioning in AWS Glue.

Known issues for AWS Glue

Note the following known issues for AWS Glue.

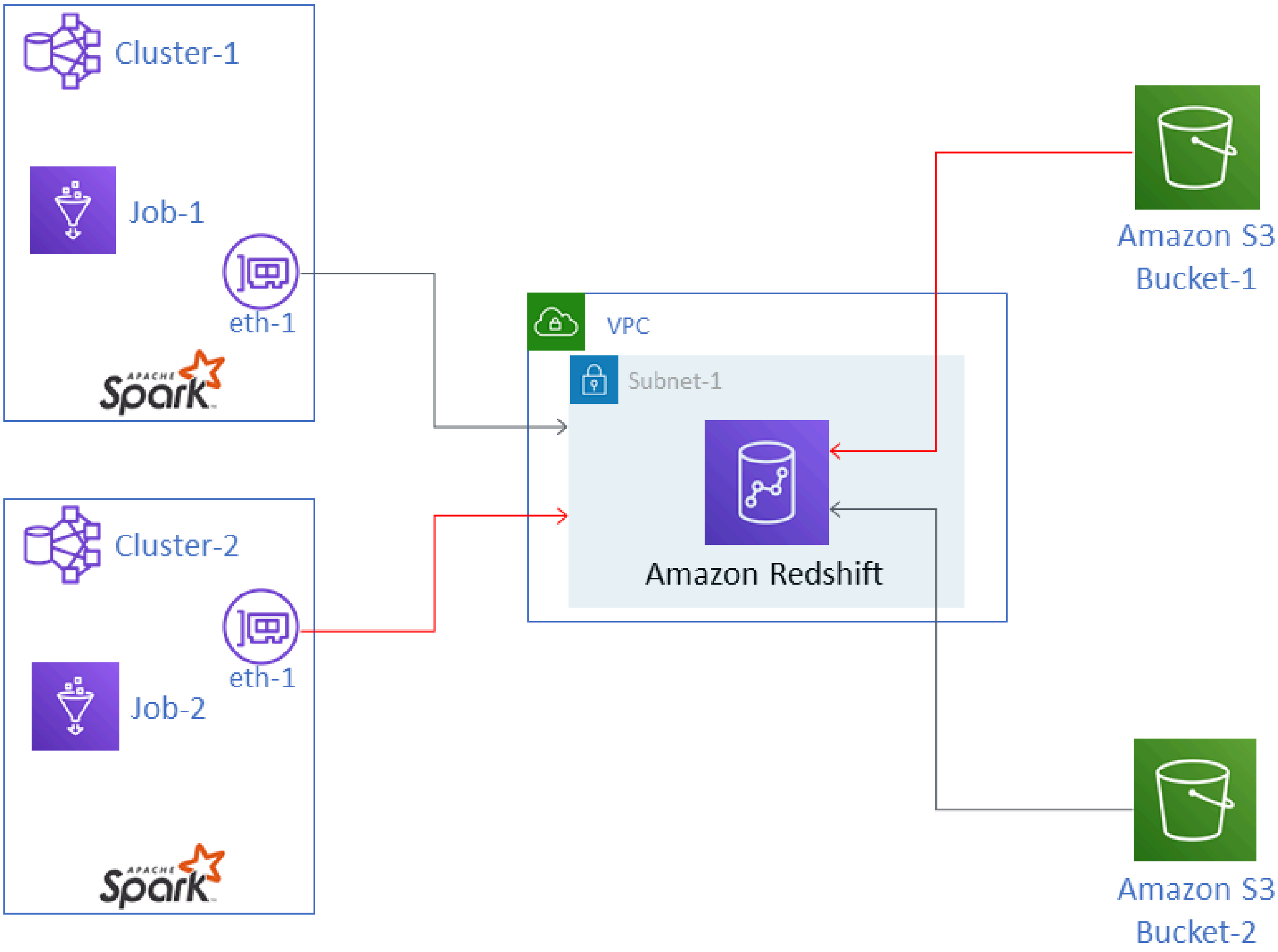
Topics

- Preventing cross-job data access

Preventing cross-job data access

Consider the situation where you have two AWS Glue Spark jobs in a single AWS Account, each running in a separate AWS Glue Spark cluster. The jobs are using AWS Glue connections to access resources in the same virtual private cloud (VPC). In this situation, a job running in one cluster might be able to access the data from the job running in the other cluster.

The following diagram illustrates an example of this situation.


Cluster-1
Job-1
eth-1
Spark
Cluster-2
eth-1
Job-2
Spark
VPC
Subnet-1
Amazon Redshift
Amazon S3
Bucket-1
Amazon S3
Bucket-2

In the diagram, AWS Glue `Job-1` is running in `Cluster-1`, and Job-2 is running in `Cluster-2`. Both jobs are working with the same instance of Amazon Redshift, which resides in `Subnet-1` of a VPC. `Subnet-1` could be a public or private subnet.

`Job-1` is transforming data from Amazon Simple Storage Service (Amazon S3) `Bucket-1` and writing the data to Amazon Redshift. `Job-2` is doing the same with data in `Bucket-2`. `Job-1` uses the AWS Identity and Access Management (IAM) role `Role-1` (not shown), which gives access to `Bucket-1`. `Job-2` uses `Role-2` (not shown), which gives access to `Bucket-2`.

These jobs have network paths that enable them to communicate with each other's clusters and thus access each other's data. For example, `Job-2` could access data in `Bucket-1`. In the diagram, this is shown as the path in red.

To prevent this situation, we recommend that you attach different security configurations to `Job-1` and `Job-2`. By attaching the security configurations, cross-job access to data is blocked by virtue of certificates that AWS Glue creates. The security configurations can be *dummy* configurations. That is, you can create the security configurations without enabling encryption of Amazon S3 data, Amazon CloudWatch data, or job bookmarks. All three encryption options can be disabled.

For information about security configurations, see the section called “Encrypting data written by AWS Glue”.

To attach a security configuration to a job

1. Open the AWS Glue console at https://console.aws.amazon.com/glue/.
2. On the **Configure the job properties** page for the job, expand the **Security configuration, script libraries, and job parameters** section.
3. Select a security configuration in the list.

Documentation history for AWS Glue

Change	Description	Date
Support for new worker types G.12X, G.16X, and memory-optimized R worker types	Added support for new worker types G.12X, G.16X, and memory-optimized R.1X, R.2X, R.4X, R.8X worker types for AWS Glue jobs. These new worker types provide additional compute and memory options for different workload requirements. For more information, see Adding jobs in AWS Glue.	January 30, 2025
Support for 14 more new SaaS native connectors for AWS Glue	Fourteen more SaaS native connectors were added to AWS Glue. For more information, see Adding an AWS Glue connection.	January 30, 2025
Support for 16 more new SaaS native connectors for AWS Glue	Sixteen more SaaS native connectors were added to AWS Glue. For more information, see Adding an AWS Glue connection.	December 17, 2024
Automatic column statistics generation	AWS Glue Data Quality now supports Amazon SageMaker AI LakeHouse tables and AWS Lake Formation managed Iceberg, Delta and HUDI tables in Data Catalog and ETL. For more information, see AWS Glue Data Quality.	December 6, 2024

Support for zero-ETL integrations	Zero-ETL is a set of fully managed integrations by AWS that minimizes the need to build ETL data pipelines. For more information, see Zero-ETL integrations.	December 3, 2024
Support for reusable connections	A new AWS Glue connections schema provides a unified way to manage data connections across AWS services and applications, such as AWS Glue, Amazon Athena, and Amazon SageMaker Unified Studio. For more information, see Connecting to data.	December 3, 2024
Support for AWS Glue version 5.0.	Added information about support for AWS Glue version 5.0. Features include an Apache Spark update to 3.52, Java update to 17, open table format updates, Spark native fine-grained access control, Sagemaker Lakehouse and data warehouse abstraction integration, support in Sagemaker Unified Studio, and more. For more information, see AWS Glue Release Notes and Migrating AWS Glue jobs to AWS Glue version 5.0.	December 3, 2024

Connecting to AWS Glue Data Catalog using AWS Glue Iceberg REST endpoint	AWS Glue's Iceberg REST endpoint supports API operations specified in the Apache Iceberg REST specification. Using an Iceberg REST client, you can connect your application running on an analytics engine to the REST catalog hosted in the Data Catalog. For more information, see Accessing the Data Catalog.	December 3, 2024
Automatic column statistics generation	Automatically generate column statistics for new tables in the AWS Glue Data Catalog. For more information, see Automatic column statistics generation.	December 3, 2024
Support for Generative AI upgrades for Apache Spark in AWS Glue	Spark Upgrades in AWS Glue enables data engineers and developers to upgrade and migrate upgrade and migrate their existing AWS Glue Spark jobs to the latest Spark releases using generative AI. For more information, see Upgrade analysis with AI.	November 22, 2024

Support for Generative AI troubleshooting for Apache Spark in AWS Glue	Generative AI Troubleshooting for Apache Spark jobs in AWS Glue helps data engineers and scientists diagnose and fix issues in their Spark applications with ease. For more information, see Troubleshooting Spark jobs with AI.	November 22, 2024
Support for Iceberg optimizers to access Amazon S3 buckets in a VPC	AWS Glue Data Catalog supports Iceberg table optimizers to access Amazon S3 buckets from a specific Virtual Private Cloud (VPC) using AWS Glue network connection. For more information, see Optimizing Iceberg tables.	November 20, 2024
Support for nine more new SaaS native connectors for AWS Glue	Nine more SaaS native connectors were added to AWS Glue. For more information, see Adding an AWS Glue connection.	November 19, 2024
Support for ten new SaaS native connectors for AWS Glue	Ten SaaS native connectors were added to AWS Glue. For more information, see Adding an AWS Glue connection.	November 15, 2024

Support for job run queuing for AWS Glue ETL jobs	You can turn on job run queuing to run jobs later, when they cannot run immediately due to service quotas. For more information, see Configuring job properties for Spark jobs in AWS Glue.	September 3, 2024
Updated policy changes	Documented changes to the **AwsGlueSessionUserRestrictedNotebookPolicy** and **AwsGlueSessionUserRestrictedNotebookServiceRole** policies, required for supporting on tag-on-create for sessions with owner tag key. For more information, see AWS Glue updates to AWS managed policies.	August 30, 2024
Anomaly Detection and Dynamic Rules are now generally available	AWS Glue Data Quality utilizes a machine learning algorithm to learn from past trends and then predict future values to detect anomalies. Dynamic Rules allows you to provide dynamic thresholds. For more information, see Optimizing query performance for Iceberg tables .	August 7, 2024

Updated policy changes	Documented changes to the **AwsGlueSessionUserRestrictedPolicy** and **AwsGlueSessionUserRestrictedServiceRole** policies, required for supporting on tag-on-create for sessions with owner tag key. For more information, see AWS Glue updates to AWS managed policies.	August 5, 2024
Generating column statistics for Iceberg tables is now generally available	AWS Glue supports calculating and updating number of distinct values (NDVs) for each column in Iceberg tables. For more information, see Anomaly detection in AWS Glue Data Quality and Dynamic rules.	July 9, 2024
Support for AWS Glue usage profiles	Admins can create AWS Glue usage profiles for various classes of users within the account, such as developers, testers, and product teams. This flexibility allows administrators to apply different usage and cost controls for each class of users. For more information, see Setting up AWS Glue usage profiles.	June 18, 2024

Support for a Salesforce connector for AWS Glue for Spark	Added information about a new AWS Glue connector for Salesforce. This feature allows you to use AWS Glue for Spark to read from and write to Salesforce in AWS Glue 4.0 and later versions. For more information, see Connecting to Salesforce.	May 22, 2024

Amazon Q data integration in AWS Glue (GA)	Amazon Q data integration in AWS Glue is a new generative AI capability of AWS Glue that enables data engineers and ETL developers to build data integration jobs using natural language. Engineers and developers can ask Q to author jobs, troubleshoot issues and answer questions about AWS Glue and data integration. For more information, see Amazon Q data integration in AWS Glue. This feature includes an update to the `AwsGlueSessionUserRestrictedPolicy` , `AwsGlueSessionUserRestrictedNotebookServiceRole` , and `AwsGlueSessionUserRestrictedServiceRole` AWS managed policies. For more information, see AWS Glue updates to AWS managed policies.	April 30, 2024

Amazon Q data integration in AWS Glue (preview)	Amazon Q data integration in AWS Glue is a new generative AI capability of AWS Glue that enables data engineers and ETL developers to build data integration jobs using natural language. Engineers and developers can ask Q to author jobs, troubleshoot issues and answer questions about AWS Glue and data integration. For more information, see Amazon Q data integration in AWS Glue. This feature includes an update to the `AwsGlueSessionUserRestrictedNotebookPolicy` AWS managed policy. For more information, see AWS Glue updates to AWS managed policies.	January 30, 2024
Update to the documentation for AWS Glue Streaming	Added a new chapter with new and reorganized content for AWS Glue Streaming. This content describes how streaming works with AWS Glue, the characteristics of real-time data processing, and how to monitor your streaming jobs. For more information, see AWS Glue Streaming.	December 27, 2023

Support for using fine-grained sensitive data detection	The Detect Sensitive Data transform provides the ability to detect, mask, or remove entities that you define, or are pre-defined by AWS Glue. Fine-grained actions further allows you to apply a specific action per entity. For more information, see Using fine-grained sensitive data detection.	November 26, 2023
Support for monitoring jobs with AWS Glue Observability metrics	Use AWS Glue Observability metrics to generate insights into what is happening inside your AWS Glue for Apache Spark jobs to improve triaging and analysis of issues. For more information, see Monitoring with AWS Glue Observability metrics.	November 26, 2023
Support for anomaly detection in AWS Glue Data Quality	AWS Glue Data Quality anomaly detection applies machine learning (ML) algorithms on data statistics over time to detect abnormal patterns and hidden data quality issues that are hard to detect through rules. For more information, see Anomaly detection in AWS Glue Data Quality.	November 26, 2023

Update to default Spark UI logging behavior	Spark jobs generating Spark UI logs will now write with a different filename pattern to support Spark UI in the AWS Glue console. This does not change CloudWatch log behavior. You can revert to the legacy behavior by updating your job configuration. For more information, see Monitoring jobs using the Apache Spark web UI.	November 17, 2023
Support for new data sources in AWS Glue for Spark	Connections to Amazon OpenSearch Service, Azure SQL, Azure Cosmos for NoSQL, SAP HANA Teradata Vantage and Vertica are now supported natively within AWS Glue. Additionally, connections to these data sources, along with MongoDB, are now available for use in the AWS Glue Studio visual editor. For more information, see Connection types and options for ETL in AWS Glue for Spark for information about AWS Glue for Spark support and Adding an AWS Glue connection for information about use in the AWS Glue Studio visual editor.	November 17, 2023

Support for generating column statistics	You can compute column-level statistics for AWS Glue Data Catalog tables in data formats such as Parquet, ORC, JSON, ION, CSV, and XML without setting up additional data pipelines. For more information, see Working with column statistics.	November 16, 2023
Support for data compaction for Iceberg tables	For better read performance by AWS analytics services such as Amazon Athena and Amazon EMR, and AWS Glue ETL jobs, Data Catalog provides managed compaction (a process that compacts small Amazon S3 objects into larger objects) for Iceberg tables in Data Catalog. For more information, see Optimizing Iceberg tables.	November 13, 2023
Update to job run wait behavior	Standard Spark and Python shell job runs will now transition to `WAITING` in certain situations, rather than immediately transitioning to `FAILED`. For more information, see AWS Glue job run statuses.	November 8, 2023

AWS Glue Studio user guide consolidated into AWS Glue developer guide	The AWS Glue Studio user guide has been moved into the developer guide to create a single unified user guide for AWS Glue Studio, the AWS Glue console, and AWS Glue Studio programmatic access.	October 25, 2023
Update to the AWSGlueServiceNotebookRole AWS managed policy	Added information about a minor update to the AWSGlueServiceNotebookRole AWS managed policy. For more information, see AWS Glue Updates to AWS Managed Policies.	October 9, 2023
AWS Glue Studio supports five new built-in transforms	AWS Glue Studio supports the following five new built-in transforms: Record matching, Remove null rows, Parse JSON column, Extract JSON path, and Regex extractor. For more information, see Editing AWS Glue managed data transform nodes.	August 11, 2023
Update to the AWSGlueServiceRole AWS managed policy	Added information about a minor update to the AWSGlueServiceRole AWS managed policy. For more information, see AWS Glue Updates to AWS Managed Policies.	August 4, 2023

Support for crawling Apache Hudi tables	Added information about using AWS Glue to crawl Hudi tables in Amazon S3 buckets and registering the Hudi tables to the AWS Glue Data Catalog. For more information, see Which data stores can I crawl?, and Crawler properties.	July 21, 2023
Update to the AWSGlueConsoleFullAccess AWS managed policy	Added information about a minor update to the AWSGlueConsoleFullAccess AWS managed policy. For more information, see AWS Glue Updates to AWS Managed Policies.	July 14, 2023
Support for crawling Apache Iceberg tables	Added information about using AWS Glue to crawl Iceberg tables in Amazon S3 buckets and registering the Iceberg tables to the AWS Glue Data Catalog. For more information, see Which data stores can I crawl?, and Crawler properties.	July 7, 2023
Support for AWS Glue with Ray	Added information about AWS Glue with Ray, a new engine that can back AWS Glue jobs. Reorganized existing AWS Glue with Spark content to disambiguate.	May 30, 2023

Support for AWS Glue Data Quality (GA)	AWS Glue Data Quality is now generally available. AWS Glue Data Quality helps you evaluate and monitor the quality of your data. For information about how to use AWS Glue Data Quality with Data Catalog, see AWS Glue Data Quality. To learn about AWS Glue Data Quality for AWS Glue Studio, see Evaluating data quality with AWS Glue Studio.	May 24, 2023
Support for larger worker types for Apache Spark jobs	Support is now available for use of the `G.4X` and `G.8X` worker types for Apache Spark jobs. These worker types are appropriate for jobs whose workloads contain your most demanding transforms, aggregations, joins, and queries. For more information, see Adding jobs in AWS Glue.	May 8, 2023
Support for creating partition indexes when crawling tables	Added information about how crawlers support the creation of partition indexes for tables that the crawler detects. For more information, see Setting the partition index crawler configuration option.	April 24, 2023

Support for resource usage metrics	Added information about viewing the service's resource usage and configuring alarms in Amazon CloudWatch. For more information, see AWS Glue resource monitoring.	April 7, 2023
Update to the AWSGlueConsoleFullAccess AWS managed policy	Added information about a minor update to the AWSGlueConsoleFullAccess AWS managed policy. For more information, see AWS Glue Updates to AWS Managed Policies.	March 28, 2023
Added guidance for using AWS Glue with an AWS SDK with examples	The AWS Glue Developer Guide has two new sections that provide information to help you use AWS Glue with an AWS SDK. For more information, see Using AWS Glue with an AWS SDK and Code examples for AWS Glue using AWS SDKs.	February 23, 2023
Update to the documentation for IAM with AWS Glue	Reorganized and added information on using IAM with AWS Glue. For more information, see Identity and access management for AWS Glue.	February 15, 2023

Support for running streaming ETL jobs in AWS Glue version 4.0	Added information about support for running streaming ETL jobs in Glue version 4.0, and new options for connecting to a Kafka cluster or an Amazon Managed Streaming for Apache Kafka cluster, and Amazon Kinesis Data Streams. For more information, see Adding Streaming ETL Jobs in AWS Glue and Connection types and options for ETL in AWS Glue.	February 8, 2023
Support for crawling MongoDB Atlas data sources	Added information about using AWS Glue to crawl MongoDB Atlas data sources. For more information, see Which data stores can I crawl?, MongoDB and MongoDB Atlas connection properties, and Using a MongoDB or MongoDB Atlas connection.	February 6, 2023

Support for crawling Delta Lake tables using a native Delta Lake connector	Added information about using AWS Glue to crawl Delta Lake tables using a native Delta Lake connector. This feature allows you to use AWS query engines to query the Delta transaction log directly and use features such as time travel and ACID guarantees, and to sync your Delta Lake metadata from Amazon S3 transaction files into the Data Catalog to enable column permissions on your queries in Lake Formation. For more information, see How to specify configuration options for a Delta Lake data store, and Querying Delta Lake tables.	December 15, 2022
Support for AWS Glue Data Quality (preview)	Support is now available for AWS Glue Data Quality (preview). AWS Glue Data Quality helps you evaluate and monitor the quality of your data when you use AWS Glue 3.0. For information about how to use AWS Glue Data Quality with Data Catalog, see AWS Glue Data Quality (preview). To learn about AWS Glue Data Quality for AWS Glue Studio, see Evaluating data quality with AWS Glue Studio.	November 30, 2022

Support for a new Amazon Redshift Spark connector with new features and performance improvements	Support is now available for a new Amazon Redshift Spark connector with a new JDBC driver for use with AWS Glue ETL jobs to build Apache Spark applications that read from and write to data in Amazon Redshift as part of your data ingestion and transformation pipelines. For more information, see Moving data to and from Amazon Redshift.	November 29, 2022
Support for AWS Glue version 4.0.	Added information about support for AWS Glue version 4.0. Features include native support for open-data lake frameworks with Apache Hudi, Delta Lake, and Apache Iceberg, and native support for the Amazon S3-based Cloud Shuffle Storage Plugin (an Apache Spark plugin) to use Amazon S3 for shuffling and elastic storage capacity. For more information, see AWS Glue Release Notes and Migrating AWS Glue jobs to AWS Glue version 4.0.	November 28, 2022

AWS Glue Studio now offers custom visual transforms	Custom visual transforms let customers define, reuse, and share business-specific ETL logic among their teams. For more information, see Custom visual transforms .	November 28, 2022
Support for using the AWS Glue crawler to publish metadata for JDBC data stores	Support is now available for using the AWS Glue crawler to publish metadata such as comments and rawtypes to the Data Catalog for JDBC data stores. For more information, see Parameters set on Data Catalog tables by crawler, Crawler properties, and JdbcTarget structure.	November 18, 2022
Support for crawling Snowflake data stores	Support is now available for using AWS Glue to crawl Snowflake tables and views, and to publish the metadata to the Data Catalog as a table entry. For Snowflake external tables in Amazon S3, the crawler also crawls the Amazon S3 location and the file format type of the external table and populates as Table parameters. For more information, see Which data stores can I crawl?, AWS Glue connection properties, and Parameters set on Data Catalog tables by crawler.	November 18, 2022

Support for improved shuffle management of your Spark applications	Support is now available for a new Cloud Shuffle Storage Plugin for Apache Spark. For more information, see AWS Glue Spark shuffle plugin with Amazon S3 and Cloud Shuffle Storage Plugin for Apache Spark.	November 15, 2022
Added support for Data Catalog targets when accelerating crawls Amazon S3 event notifications	In addition to the existing support for Amazon S3 targets, support is now available for accelerating crawls for Data Catalog targets using Amazon S3 event notifications. For more information, see Accelerating Crawls Using Amazon S3 Event Notifications.	October 13, 2022
Support for specifying the maximum number of tables a crawler can create	Support is now available for specifying the maximum number of tables the crawler is allowed to create. For more information, see How to specify the maximum number of tables the crawler is allowed to create.	September 6, 2022
Support for Python 3.9 in Python shell jobs in AWS Glue	Support is now available for running scripts compatible with Python 3.9 in Python shell jobs in AWS Glue, and for choosing to use pre-packaged library sets. For more information, see Python shell jobs in AWS Glue.	August 11, 2022

Support for running non-urgent or non-time sensitive AWS Glue jobs on spare capacity	Support is now available for the configuration of flexible job runs for non-urgent jobs such as pre-production jobs, testing, and one-time data loads. For more information, see Adding jobs in AWS Glue.	August 9, 2022
Support for a new worker type for streaming jobs	Support is now available for use of the `G.025X` worker type for low volume streaming jobs. For more information, see Adding jobs in AWS Glue.	July 14, 2022
Support for the use of Kafka SASL in AWS Glue connections	Support is now available for use of Kafka SASL in AWS Glue connections. For more information, see AWS Glue Kafka connection properties for client authentication.	July 5, 2022
Support for Apache kafka connector for protobuf schemas	Support is now available for Apache Kafka Connector for Protobuf schemas. For more information, see AWS Glue Schema Registry.	June 9, 2022
Support for Auto Scaling for AWS Glue jobs (GA)	Added information on using Auto Scaling for jobs in AWS Glue version 3.0 to dynamically scale compute resources. For more information, see Using Auto Scaling for AWS Glue.	April 14, 2022

Update to the documentation for AWS Glue developing and testing AWS Glue job scripts	Reorganized and added information on the available development and testing methods for AWS Glue, including instructions for developing with Docker. For more information, see Developing and testing AWS Glue job scripts.	March 14, 2022
Addition of protocol buffers (protobuf) as a supported data format for the AWS Glue schema registry	Added information about Protobuf as a supported data format (in addition to AVRO and JSON). For more information, see AWS Glue Schema Registry.	February 25, 2022
Support for crawling Delta Lake tables	Added information about using AWS Glue to crawl Delta Lake tables. For more information, see How to specify configuration options for a Delta Lake data store.	February 24, 2022
Support for AWS Glue job insights	Added information about using AWS Glue job insights to simplify job debugging and optimization for your AWS Glue jobs. For more information, see Monitoring with AWS Glue job insights.	February 8, 2022

Support for crawling Amazon S3 backed Data Catalog tables using a VPC endpoint	In addition to Amazon S3 data stores, you can configure your Amazon S3 backed Data Catalog tables to be accessed only by an Amazon Virtual Private Cloud environment (Amazon VPC), for security, auditing, or control purposes. For more information, see Crawling an Amazon S3 Data Store or Amazon S3 backed Data Catalog tables using a VPC Endpoint.	February 3, 2022
Support for Lake Formation governed tables	Added information about AWS Glue support for Lake Formation governed tables, which support ACID transactions, automatic data compaction, and time-travel queries. For more information, see AWS Glue API and the AWS Lake Formation developer guide.	November 30, 2021
New AWS managed policies added for interactive sessions and notebooks	New managed policies for IAM provided enhanced security for using AWS Glue with interactive sessions and notebooks. For more information, see AWS Managed Policies for AWS Glue.	November 30, 2021

Glue schema registry now supported with streaming jobs	You can create streaming jobs that access tables that are part of the Glue Schema Registry. For more information see AWS Glue Schema Registry and Adding Streaming ETL Jobs in AWS Glue.	November 15, 2021
Support for new machine learning features	Added information about new features for the Find matches machine learning transform , including incremental matching and match scoring. For more information, see Finding Incremental Matches and Estimating the Quality of Matches using Match Confidence Scores.	October 31, 2021
(Private preview) Support for AWS Glue flex jobs	Added information about configuring AWS Glue Spark jobs with a flexible execution class, appropriate for time-insensitive jobs whose start and completion times may vary. For more information, see Adding Jobs in AWS Glue.	October 29, 2021
Support for accelerating crawls using Amazon S3 event notifications	Added information about accelerating crawls using Amazon S3 event notifications. For more information, see Accelerating Crawls Using Amazon S3 Event Notifications.	October 15, 2021

Additional security configuration options related to access-control and VPCs	Added information about how you can configure new access control permissions on AWS Glue and configuration of VPCs. For more information, see AWS Tags in AWS Glue, Identity-Based Policies (IAM Policies) that Control Settings Using Condition Keys or Context Keys, and Configuring all AWS calls to go through your VPC.	October 13, 2021
Support for VPC endpoint policies	Added information about support for Virtual Private Cloud (VPC) endpoint policies in AWS Glue. For more information, see AWS Glue and interface VPC endpoints (AWS PrivateLink).	October 11, 2021
Glue Studio is now available in China	AWS Glue Studio is now available in the China Beijing and Ningxia regions.	October 11, 2021
AWS Glue Studio offers notebook authoring, for interactive job editing	Notebooks help you to write and execute code, visualize the results, and share insights. Typically, data scientists use notebooks for experiments and data exploration tasks. For more information, see Using Notebooks.	October 1, 2021

Direct access to streaming sources now available	When adding data sources to your ETL job in the visual editor, you can supply information to access the data stream instead of having to use a Data Catalog database and table.	September 30, 2021
Documented the AWS Glue version support policy	Added information about the AWS Glue version support policy and the end of life phases for certain AWS Glue versions. For more information, see AWS Glue version support policy.	September 24, 2021
Custom connectors can now be used with data previews	When editing data source node using a custom connector, you can preview the dataset by choosing the Dat preview tab. For more information, see Custom Connectors .	September 24, 2021

Support for AWS Glue interactive sessions (private preview)	(Private preview) Added information about using AWS Glue interactive sessions to run Spark workloads in the cloud from any Jupyter Notebook. Interactive sessions are the preferred method for developing your AWS Glue extract, transform , and load (ETL) code when you use AWS Glue 2.0 or later. For more information, see Setting Up and Running AWS Glue interactive sessions for Jupyter Notebook.	August 24, 2021
Support for creating workflows from blueprints (GA)	Added information about coding common extract, transform, and load (ETL) use cases in blueprints and then creating workflows from blueprints. Enables data analysts to easily create and run complex ETL processes . For more information, see Performing Complex ETL Activities Using blueprints and Workflows in AWS Glue.	August 23, 2021

Support for AWS Glue version 3.0.	Added information about support for AWS Glue version 3.0 which supports the Apache Spark 3.0 engine upgrade for running Apache Spark ETL jobs, and other optimizations and upgrades. For more information, see AWS Glue Release Notes and Migrating AWS Glue jobs to AWS Glue version 3.0. Other features in this release include the AWS Glue shuffle manager, a SIMD vectorized CSV reader, and catalog partition predicates. For more information see AWS Glue Spark shuffle manager with Amazon S3, Format Options for ETL Inputs and Outputs in AWS Glue, and Server-side filtering using catalog partition predicates.	August 18, 2021
AWS GovCloud (US) Region	AWS Glue Studio is now available in the AWS GovCloud (US) Region	August 18, 2021
Python shell authoring available in AWS Glue Studio	When creating a new job, you can now choose to create a Python shell job. For more information, see Start the job creation process and Editing Python shell jobs in AWS Glue Studio.	August 13, 2021

Support for starting a workflow with an Amazon EventBridge event	Added information about how AWS Glue can be an event consumer in an event-driven architecture. For more information, see Starting an AWS Glue Workflow with an Amazon EventBridge Event and Viewing the EventBridge Events That Started a Workflow.	July 14, 2021
Addition of JSON as a supported data format for the AWS Glue schema registry	Added information about JSON as a supported data format (in addition to AVRO). For more information, see AWS Glue Schema Registry.	June 30, 2021
Create AWS Glue streaming jobs without a Data Catalog table	The `create_data_frame_from_options` Python function or `getSource` for Scala scripts support creating streaming ETL jobs that reference the data streams directly instead of requiring a Data Catalog table.	June 15, 2021
AWS Glue machine learning transforms now support AWS Key Management Service keys	You can specify a security configuration or AWS KMS key when configuring AWS Glue Machine Learning transforms with the console, the CLI, or the AWS Glue APIs. For more information, see Using Data Encryption with Machine Learning Transforms and AWS Glue Machine Learning API.	June 15, 2021

Update to the AWSGlueConsoleFullAccess AWS managed policy	Added information about a minor update to the AWSGlueConsoleFullAccess AWS managed policy. For more information, see AWS Glue Updates to AWS Managed Policies.	June 10, 2021
View your job's dataset while creating and editing jobs	You can use the new **Data preview** tab for a node in your job diagram to see a sample of the data processed by that node. For more information, see Using data previews in the visual job editor.	June 7, 2021
Support for specifying a value that indicates the table location for the crawler output.	Added information about specifying a value that indicates the table location when configuring the crawler's output. For more information, see How to specify the table location.	June 4, 2021
Support for crawling a sample of files in a dataset when crawling an Amazon S3 data store	Added information about how to crawl a sample of files when crawling Amazon S3. For more information, see Crawler Properties.	May 10, 2021

Support for the AWS Glue optimized parquet writer	Added information about using the AWS Glue optimized parquet writer for DynamicFrames to create or update tables with the `parquet` classification. For more information, see Creating Tables, Updating Schema, and Adding New Partitions in the Data Catalog from AWS Glue ETL Jobs and Format Options for ETL Inputs and Outputs in AWS Glue.	May 4, 2021
Support for kafka client authentication passwords	Added information about how streaming ETL jobs in AWS Glue support SSL client certificate authentication with Apache Kafka stream producers. You can now provide a custom certificate while defining an AWS Glue connection to an Apache Kafka cluster, which AWS Glue will use when authenticating with it. For more information, see AWS Glue Connection Properties and Connection API.	April 28, 2021

Support for consuming data from Amazon Kinesis Data Streams in another account in streaming ETL jobs	Added information about to create a streaming ETL job to consume data from Amazon Kinesis Data Streams in another account. For more information, see Adding Streaming ETL Jobs in AWS Glue.	March 30, 2021
SQL transform available	You can use a **SQL** transform node to write your own transform in the form of a SQL query. For more information, see Using a SQL query to transform data.	March 23, 2021
Support for creating workflows from blueprints (public preview)	(Public preview) Added information about coding common extract, transform , and load (ETL) use cases in blueprints and then creating workflows from blueprint s. Enables data analysts to easily create and run complex ETL processes. For more information, see Performing Complex ETL Activities Using blueprints and Workflows in AWS Glue.	March 22, 2021
Connectors can be used for data targets	Using a custom or AWS Marketplace connector for your data target is now supported. For more information, see Authoring jobs with custom connectors.	March 15, 2021

Support for column importance metrics for AWS Glue machine learning transforms	Added information about viewing column importance metrics when working with AWS Glue machine learning transforms. For more information see Working with Machine Learning Transforms on the AWS Glue Console	February 5, 2021
Job scheduling now available in AWS Glue Studio	You can define a time-based schedule for your job runs in AWS Glue Studio. You can use the console to create a basic schedule, or define a more complex schedule using the Unix-like cron syntax. For more information, see Schedule job runs.	December 21, 2020
AWS Glue Custom Connectors released	AWS Glue Custom Connectors allow you to discover and subscribe to connectors in AWS Marketplace. We also released AWS Glue Spark runtime interfaces to plug in connectors built for Apache Spark Datasource, Athena federated query, and JDBC APIs. For more information, see Using Connectors and connections with AWS Glue Studio.	December 21, 2020

Support for running streaming ETL jobs in AWS Glue version 2.0	Added information about support for running streaming ETL jobs in Glue version 2.0. For more information, see Adding Streaming ETL Jobs in AWS Glue.	December 18, 2020
Support for workload partitioning with bounded execution	Added information about enabling workload partitioning to configure the upper bounds on the dataset size, or the number of files processed on ETL job runs. For more information, see Workload Partitioning with Bounded Execution.	November 23, 2020
Support for enhanced partition management	Added information about how to use new APIs to add or delete a partition index to/from an existing table. For more information, see Working with Partition Indexes.	November 23, 2020
Support for the AWS Glue schema registry	Added information about using the AWS Glue Schema Registry to centrally discover, control, and evolve schemas. For more information, seeAWS Glue Schema Registry.	November 19, 2020

Support for the grok input format in streaming ETL jobs	Added information about applying Grok patterns to streaming sources such as log files. For more information, see Applying Grok Patterns to Streaming Sources.	November 17, 2020
Support for adding tags to workflows on the AWS Glue console	Added information about adding tags when creating a workflow using the AWS Glue console. For more information, see Creating and Building Out a Workflow Using the AWS Glue Console.	October 27, 2020
Support for incremental crawler runs	Added information about support for incremental crawler runs, which crawl only Amazon S3 folders added since the last run. For more information, see Incremental Crawls.	October 21, 2020

Support for schema detection for streaming ETL data sources. support for Avro streaming ETL data sources and self-managed kafka	Streaming extract, transform, and load (ETL) jobs in AWS Glue can now automatically detect the schema of incoming records and handle schema changes on a per-record basis. Self-managed Kafka data sources are now supported. Streaming ETL jobs now support the Avro format in data sources. For more information, see Streaming ETL in AWS Glue, Defining Job Properties for a Streaming ETL Job, and Notes and Restrictions for Avro Streaming Sources.	October 7, 2020
Support for crawling MongoDB and DocumentDB data sources	Added information about support for crawling MongoDB and Amazon DocumentDB (with MongoDB compatibility) data sources. For more information, see Defining Crawlers.	October 5, 2020
Support for FIPS compliance	Added information about FIPS endpoints for customers who require FIPS 140-2 validated cryptographic modules when accessing data using AWS Glue. For more information, see FIPS Compliance.	September 23, 2020

AWS Glue Studio provides an easy to use visual interface for creating and monitoring jobs	You can now use a simple graph-based interface to compose jobs that move and transform data and run them on AWS Glue. You can then use the job run dashboard in AWS Glue Studio to monitor ETL execution and ensure that your jobs are operating as intended. For more information, see AWS Glue Studio User Guide.	September 23, 2020
Support for creating table indexes to improve query performance	Added information about creating table indexes to allow you to retrieve a subset of the partitions from a table. For more information, see Working with Partition Indexes.	September 9, 2020
Support for reduced startup times when running Apache Spark ETL jobs in AWS Glue version 2.0.	Added information about support for AWS Glue version 2.0 which provides an upgraded infrastructure for running Apache Spark ETL jobs with reduced startup times, changes in logging, and support for specifying additional Python modules at the job level. For more information, see AWS Glue Release Notes and Running Spark ETL Jobs with Reduced Startup Times.	August 10, 2020

Support for limiting the number of concurrent workflow runs.	Added information about how to limit the number of concurrent workflow runs for a particular workflow. For more information, see Creating and Building Out a Workflow Using the AWS Glue Console.	August 10, 2020
Support for crawling an Amazon S3 data store using a VPC endpoint	Added information about configuring your Amazon S3 data store to be accessed only by an Amazon Virtual Private Cloud environment (Amazon VPC), for security, auditing, or control purposes. For more information, see Crawling an Amazon S3 Data Store using a VPC Endpoint.	August 7, 2020
Support for resuming workflow runs	Added information about how to resume workflow runs that only partially completed because one or more nodes (jobs or crawlers) did not complete successfully. For more information, see Repairing and Resuming a Workflow Run.	July 27, 2020

Support for enabling private CA certificates in kafka connections in AWS Glue.	Added information about new connection options that support enabling private CA certificates for Kafka connections in AWS Glue. For more information, see Connection Types and Options for ETL in AWS Glue and Special Parameters Used by AWS Glue.	July 20, 2020
Support for reading DynamoDB data in another account	Added information about AWS Glue support for reading data from another AWS account's DynamoDB table For more information, see Reading from DynamoDB Data in Another Account.	July 17, 2020
Support for a DynamoDB writer connection in AWS Glue version 1.0 or later	Added information about support for DynamoDB writer, and new or updated connection options for DynamoDB to read or write. For more information, see Connection Types and Options for ETL in AWS Glue.	July 17, 2020

Support for resource links and for cross-account access control using both AWS Glue and Lake Formation	Added content about new Data Catalog objects called resource links, and about how to manage sharing Data Catalog resources across accounts with both AWS Glue and AWS Lake Formation. For more information, see Granting Cross-Account Access and Table Resource Links.	July 7, 2020
Support for sampling records when crawling DynamoDB data stores	Added information about new properties that you can configure when crawling a DynamoDB data store. For more information, see Crawler Properties.	June 12, 2020
Support for stopping a workflow run.	Added information about how to stop a workflow run for a particular workflow. For more information, see Stopping a Workflow Run.	May 14, 2020
Support for Spark streaming ETL jobs	Added information about creating extract, transform , and load (ETL) jobs with streaming data sources. For more information, see Adding Streaming ETL Jobs in AWS Glue.	April 27, 2020

Support for creating tables, updating the schema, and adding new partitions in the Data Catalog after running an ETL job	Added information about how you can enable creating tables, updating the schema, and adding new partitions to see the results of your ETL job in the Data Catalog. For more information, see Creating Tables, Updating Schema, and Adding New Partitions in the Data Catalog from AWS Glue ETL Jobs.	April 2, 2020
Support for specifying a version for the Apache Avro data format as an ETL input and output in AWS Glue	Added information about specifying a version for the Apache Avro data format as an ETL input and output in AWS Glue. The default version 1.7. You can use the `version` format option to specify Avro version 1.8 to enable logical reading/writing. For more information, see Format Options for ETL Inputs and Outputs in AWS Glue.	March 31, 2020
Support for the EMRFS S3-optimized committer for writing Parquet data into Amazon S3	Added information about how to set a new flag to enable the EMRFR S3-optimized committer for writing Parquet data into Amazon S3 when creating or updating an AWS Glue job. For more informati on, see Special Parameters Used by AWS Glue.	March 30, 2020

Support for machine learning transforms as a resource managed by AWS resource tags	Added information about using AWS resource tags to manage and control access to your machine learning transforms in AWS Glue. You can assign AWS resource tags to jobs, triggers, endpoints , crawlers, and machine learning transforms in AWS Glue. For more information, see AWS Tags in AWS Glue.	March 2, 2020
Support for non-overrideable job arguments	Added information about support for special job parameters that cannot be overridden in triggers or when you run the job. For more information see Adding Jobs in AWS Glue.	February 12, 2020
Support for new transforms to work with datasets in Amazon S3	Added information about new transforms (Merge, Purge, and Transition) and Amazon S3 storage class exclusions for Apache Spark applications to work with datasets in Amazon S3. For more information on support for these transforms for Python, see mergeDynamicFrame and Working with Datasets in Amazon S3. For Scala, see mergeDynamicFrames and AWS Glue Scala GlueContext APIs.	January 16, 2020

Support for updating the Data Catalog with new partition information from an ETL job	Added information about how to code an extract, transform, and load (ETL) script to update the AWS Glue Data Catalog with new partition information. With this capability, you no longer have to rerun the crawler after job completion to view the new partitions. For more information see Updating the Data Catalog with New Partitions.	January 15, 2020
New tutorial: Using an SageMaker AI notebook	Added a tutorial that demonstrates how to use an Amazon SageMaker notebook to help develop your ETL and machine learning scripts. See Tutorial: Use an Amazon SageMaker Notebook with Your Development Endpoint.	January 3, 2020
Support for reading from MongoDB and Amazon DocumentDB (with MongoDB compatibility)	Added information about new connection types and connection options for reading from and writing to MongoDB and Amazon DocumentDB (with MongoDB Compatibility). For more information, see Connection Types and Options for ETL in AWS Glue.	December 17, 2019

Various corrections and clarifications	Added corrections and clarifications throughout. Removed entries from the Known Issues chapter. Added warnings that AWS Glue supports only symmetrical customer master keys (CMKs) when specifying Data Catalog encryption settings and creating security configurations. Added a note that AWS Glue does not support writing to Amazon DynamoDB.	December 9, 2019
Support for custom JDBC drivers	Added information about connecting to data sources and targets with JDBC drivers that AWS Glue does not natively support, such as MySQL version 8 and Oracle Database version 18. For more information see JDBC connectionType Values.	November 25, 2019

Support for connecting SageMaker AI notebooks to different development endpoints	Added information about how you can connect an SageMaker AI notebook to different development endpoints. Updates to describe the new console action for switching to a new development endpoint, and the new SageMaker AI IAM policy. For more information, see Working with Notebooks on the AWS Glue Console and Create an IAM Policy for Amazon SageMaker AI Notebooks.	November 21, 2019
Support for AWS Glue version in machine learning transforms	Added information about defining the AWS Glue version in a machine learning transform to indicate the which version of AWS Glue a machine learning transform is compatible with. For more information see Working with Machine Learning Transforms on the AWS Glue Console.	November 21, 2019

Support for rewinding your job bookmarks	Added information about rewinding your job bookmarks to any previous job run, resulting in the subsequent job run reprocessing data only from the bookmarked job run. Described two new sub-options for the `job-bookmark-pause` option that allow you to run a job between two bookmarks. For more information, see Tracking Processed Data Using Job Bookmarks and Special Parameters Used by AWS Glue.	October 22, 2019
Support for custom JDBC certificates for connecting to a data store	Added information about AWS Glue support of custom JDBC certificates for SSL connections to AWS Glue data sources or targets. For more information, see Working with Connections on the AWS Glue Console.	October 10, 2019
Support for Python wheel	Added information about AWS Glue support of wheel files (along with egg files) as dependencies for Python shell jobs. For more information, see Providing Your Own Python Library.	September 26, 2019

Support for versioning of development endpoints in AWS Glue	Added information about defining the `Glue version` in development endpoints. `Glue version` determines the versions of Apache Spark and Python that AWS Glue supports. For more information, see Adding a Development Endpoint.	September 19, 2019
Support for monitoring AWS Glue using Spark UI	Added information about using Apache Spark UI to monitor and debug AWS Glue ETL jobs running on the AWS Glue job system, and Spark applications on AWS Glue development endpoints. For more information, see Monitoring AWS Glue Using Spark UI.	September 19, 2019
Enhancement of support for local ETL script development using the public AWS Glue ETL library	Updated the AWS Glue ETL library content to reflect that AWS Glue version 1.0 is now supported. For more information, see Developing and Testing ETL Scripts Locally Using the AWS Glue ETL Library.	September 18, 2019

Support for excluding Amazon S3 storage classes when running jobs	Added information about excluding Amazon S3 storage classes when running AWS Glue ETL jobs that read files or partitions from Amazon S3. For more information, see Excluding Amazon S3 Storage Classes.	August 29, 2019
Support for local ETL script development using the public AWS Glue ETL library	Added information about how to develop and test Python and Scala ETL scripts locally without the need for a network connection. For more information, see Developing and Testing ETL Scripts Locally Using the AWS Glue ETL Library.	August 28, 2019
Known issues	Added information about known issues in AWS Glue. For more information, see Known Issues for AWS Glue.	August 28, 2019
Support for machine learning transforms in AWS Glue	Added information about machine learning capabilities provided by AWS Glue to create custom transforms. You can create these transforms when you create a job. For more information, see Machine Learning Transforms in AWS Glue.	August 8, 2019

Support for shared Amazon Virtual Private Cloud	Added information about AWS Glue support for shared Amazon Virtual Private Cloud. For more information, see Shared Amazon VPCs.	August 6, 2019
Support for versioning in AWS Glue	Added information about defining the `Glue version` in job properties. AWS Glue version determines the versions of Apache Spark and Python that AWS Glue supports. For more information, see Adding Jobs in AWS Glue.	July 24, 2019
Support for additional configuration options for development endpoints	Added information about configuration options for development endpoints that have memory-intensive workloads. You can choose from two new configurations that provide more memory per executor. For more information, see Working with Development Endpoints on the AWS Glue Console.	July 24, 2019

Support for performing extract, transfer, and load (ETL) activities using workflows	Added information about using a new construct called a workflow to design a complex multi-job extract, transform , and load (ETL) activity that AWS Glue can run and track as a single entity. For more information, see Performing Complex ETL Activities Using Workflows in AWS Glue.	June 20, 2019
Support for Python 3.6 in Python shell jobs	Added information about support for Python 3.6 in Python shell jobs. You can specify either Python 2.7 or Python 3.6 as a job property. For more information, see Adding Python Shell Jobs inAWS Glue.	June 5, 2019
Support for virtual private cloud (VPC) endpoints	Added information about connecting directly to AWS Glue through an interface endpoint in your VPC. When you use a VPC interface endpoint, communication between your VPC and AWS Glue is conducted entirely and securely within the AWS network. For more information, see Using AWS Glue with VPC Endpoints.	June 4, 2019

Support for real-time, continuous logging for AWS Glue jobs.	Added information about enabling and viewing real-time Apache Spark job logs in CloudWatch including the driver logs, each of the executor logs, and a Spark job progress bar. For more information, see Continuous Logging for AWS Glue Jobs.	May 28, 2019
Support for existing Data Catalog tables as crawler sources	Added information about specifying a list of existing Data Catalog tables as crawler sources. Crawlers can then detect changes to table schemas, update table definitions, and register new partitions as new data becomes available. For more information, see Crawler Properties.	May 10, 2019
Support for additional configuration options for memory-intensive jobs	Added information about configuration options for Apache Spark jobs with memory-intensive workloads. You can choose from two new configurations that provide more memory per executor. For more information, see Adding Jobs in AWS Glue.	April 5, 2019

Support for CSV custom classifiers	Added information about using a custom CSV classifier to infer the schema of various types of CSV data. For more information, see Writing Custom Classifiers.	March 26, 2019
Support for AWS resource tags	Added information about using AWS resource tags to help you manage and control access to your AWS Glue resources. You can assign AWS resource tags to jobs, triggers, endpoints, and crawlers in AWS Glue. For more information, see AWS Tags in AWS Glue.	March 20, 2019
Support of Data Catalog for Spark SQL jobs	Added information about configuring your AWS Glue jobs and development endpoints to use the AWS Glue Data Catalog as an external Apache Hive Metastore. This allows jobs and development endpoints to directly run Apache Spark SQL queries against the tables stored in the AWS Glue Data Catalog. For more information, see AWS Glue Data Catalog Support for Spark SQL Jobs.	March 14, 2019

Support for Python shell jobs	Added information about Python shell jobs and the new field **Maximum capacity**. For more information, see Adding Python Shell Jobs in AWS Glue.	January 18, 2019
Support for notifications when there are changes to databases and tables	Added information about events that are generated for changes to database, table, and partition API calls. You can configure actions in CloudWatch Events to respond to these events. For more information, see Automating AWS Glue with CloudWatch Events.	January 16, 2019
Support for encrypting connection passwords	Added information about encrypting passwords used in connection objects. For more information, see Encrypting Connection Passwords.	December 11, 2018
Support for resource-level permission and resource-based policies	Added information about using resource-level permissions and resource-based policies with AWS Glue. For more information, see the topics within Security in AWS Glue.	October 15, 2018

Support for SageMaker AI notebooks	Added information about using SageMaker AI notebooks with AWS Glue development endpoints. For more information, see Managing Notebooks.	October 5, 2018
Support for encryption	Added information about using encryption with AWS Glue. For more information, see Encryption at Rest, Encryption in Transit, and Setting Up Encryption in AWS Glue.	August 24, 2018
Support for Apache Spark job metrics	Added information about the use of Apache Spark metrics for better debugging and profiling of ETL jobs. You can easily track runtime metrics such as bytes read and written, memory usage and CPU load of the driver and executors, and data shuffles among executors from the AWS Glue console. For more information, see Monitoring AWS Glue Using CloudWatch Metrics, Job Monitoring and Debugging, and Working with Jobs on the AWS Glue Console.	July 13, 2018

Support of DynamoDB as a data source	Added information about crawling DynamoDB and using it as a data source of ETL jobs. For more information, see Cataloging Tables with a Crawler and Connection Parameters.	July 10, 2018
Updates to create notebook server procedure	Updated information about how to create a notebook server on an Amazon EC2 instance associated with a development endpoint. For more information, see Creating a Notebook Server Associated with a Development Endpoint.	July 9, 2018
Updates now available over RSS	You can now subscribe to an RSS feed to receive notifications about updates to the *AWS Glue Developer Guide*.	June 25, 2018
Support delay notifications for jobs	Added information about configuring a delay threshold when a job runs. For more information, see Adding Jobs in AWS Glue.	May 25, 2018
Configure a crawler to append new columns	Added information about new configuration option for crawlers, MergeNewColumns. For more information, see Configuring a Crawler.	May 7, 2018

Support timeout of jobs	Added information about setting a timeout threshold when a job runs. For more information, see Adding Jobs in AWS Glue.	April 10, 2018
Support Scala ETL script and trigger jobs based on additional run states	Added information about using Scala as the ETL programming language. In addition, the trigger API now supports firing when any conditions are met (in addition to all conditions). Also, jobs can be triggered based on a "failed" or "stopped" job run (in addition to a "succeeded" job run).	January 12, 2018

Earlier updates

The following table describes the important changes in each release of the *AWS Glue Developer Guide* before January 2018.

Change	Description	Date
Support XML data sources and new crawler configuration option	Added information about classifying XML data sources and new crawler option for partition changes.	November 16, 2017
New transforms, support for additional Amazon RDS database engines, and developme	Added information about the map and filter transforms, support for Amazon RDS Microsoft SQL Server, and Amazon RDS Oracle, and new features for development endpoints.	September 29, 2017

Change	Description	Date
nt endpoint enhancements		
AWS Glue initial release	This is the initial release of the *AWS Glue Developer Guide*.	August 14, 2017

AWS Glossary

For the latest AWS terminology, see the AWS glossary in the *AWS Glossary Reference*.