Amazon SageMaker lakehouse architecture User Guide

lakehouse architecture

lakehouse architecture: Amazon SageMaker lakehouse architecture User Guide

Table of Contents

What is the lakehouse architecture of Amazon SageMaker?

The lakehouse architecture of Amazon SageMaker is a unified data architecture built on AWS's cloud-native infrastructure that bridges Amazon S3 data lakes and Amazon Redshift data warehouses into a cohesive analytics platform. The architecture leverages Apache Iceberg table format for cross-service interoperability and implements a shared metadata catalog that provides consistent data access patterns across storage systems.

This integrated approach enables organizations to perform analytics, machine learning, and AI workloads on a single data foundation without data movement or duplication. The architecture integrates with AWS machine learning and analytics services, enabling data scientists, analysts, and engineers to collaborate on the same datasets using their preferred tools and interfaces.

What is a data lakehouse?

A data lakehouse is an architectural pattern that unifies the scalability and cost-effectiveness of data lakes with the performance and reliability characteristics of data warehouses. This approach eliminates the traditional trade-offs between storing diverse data types and maintaining query performance for analytical workloads.

The lakehouse architecture addresses the following key limitations of isolated systems:

- **Transactional consistency** – ACID compliance ensures reliable concurrent operations
- **Schema management** – Flexible schema evolution without breaking existing queries
- **Multi-format support** – Native handling of structured, semi-structured, and unstructured data
- **Compute-storage separation** – Independent scaling of processing and storage resources
- **Open standards** – Vendor-neutral formats preventing data lock-in
- **Single source of truth** – Eliminates data silos and redundant storage costs
- **Real-time and batch processing** – Supports both streaming and historical analytics
- **Direct file access** – Enables both SQL queries and programmatic data access
- **Unified governance** – Consistent security and compliance across all data types

This architecture enables organizations to support business intelligence, advanced analytics, and machine learning workloads on the same data platform, reducing complexity and operational overhead while maintaining performance requirements for each use case.

Key Capabilities

The lakehouse architecture of Amazon SageMaker provides the following key capabilities:

- **Unified data access** – Query and access data across Amazon S3 data lakes, Amazon Redshift data warehouses, and other sources using Apache Iceberg compatible tools and engines. This includes AWS services such as Amazon Athena, Amazon Redshift, Amazon EMR, Amazon SageMaker AI, as well as third-party engines, all of which you can use to query your data in-place.
- **Integrated access control** – Fine-grained access control to your data with permissions that you can define and consistently apply across all analytics and ML tools and engines, regardless of the underlying storage formats or query engines used.
- **Open source compatibility** – Leverages open-source Apache Iceberg, enabling data interoperability across various Apache Iceberg compatible query engines and tools. This gives you the flexibility to choose your preferred tools and engines.

Key components of the lakehouse architecture of Amazon SageMaker

The lakehouse architecture has the following key components.

Storage

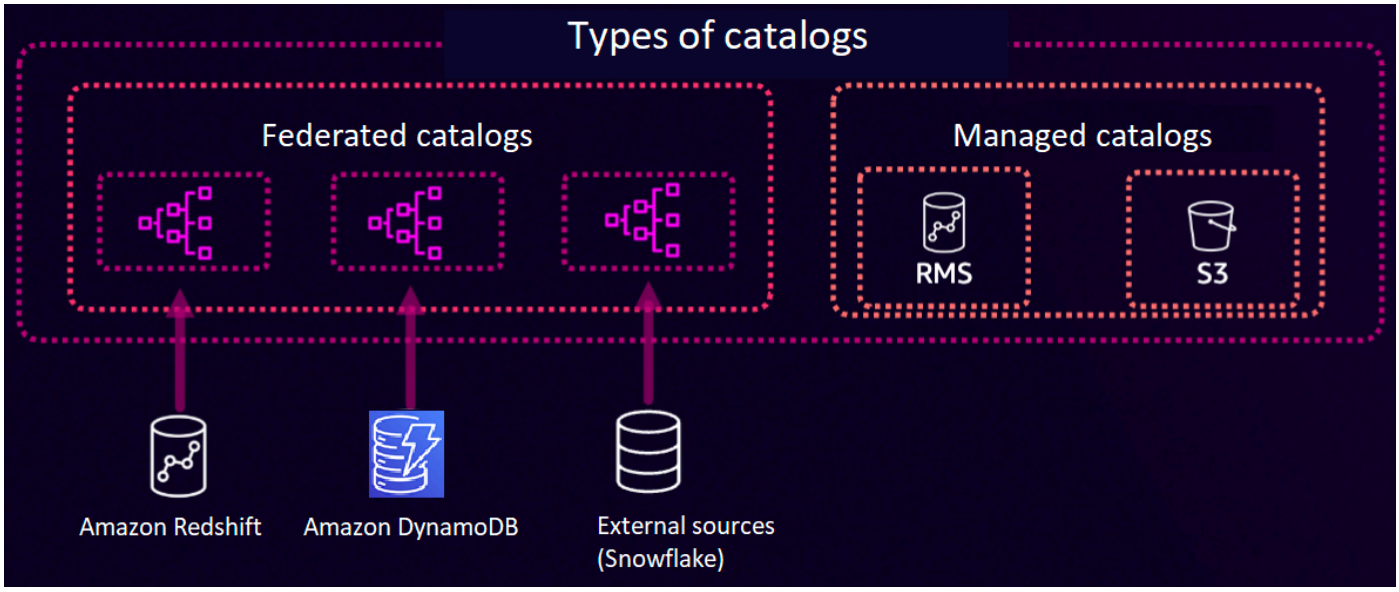
You can read and write data into Amazon S3 or Redshift Managed Storage (RMS) based on the storage type you choose to store data in the lakehouse.

Catalog

A catalog is a logical container that organizes objects from a data store, such as schemas, tables, views, or materialized views such as from Amazon Redshift. You can create nested catalogs to mirror the hierarchical structure of your data sources within SageMaker AI Lakehouse.

There are two types of catalogs in Lakehouse: federated catalogs and managed catalogs. A federated catalog mounts existing data sources you add to Lakehouse. A federated catalog

can bring existing data in data sources such as Amazon Redshift, Amazon DynamoDB, and Snowflake. A managed catalog refers to a new catalog you create using Lakehouse. A managed catalog manages data using RMS or S3, as shown in the following diagram.



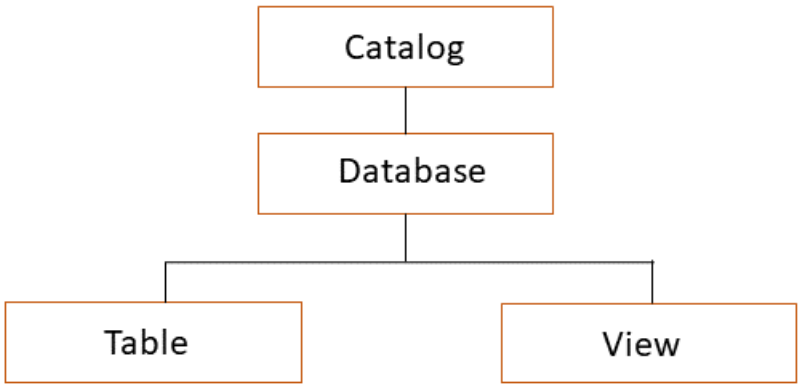
Database

Databases organize metadata tables in a catalog in the lakehouse architecture.

Table/View

Tables and views are database objects that define how to access and represent the underlying data. They specify details such as schema, partitions, storage location, storage format, and the SQL query required to access the data.

The following is a diagram of how catalogs, databases, tables/views work in Lakehouse.



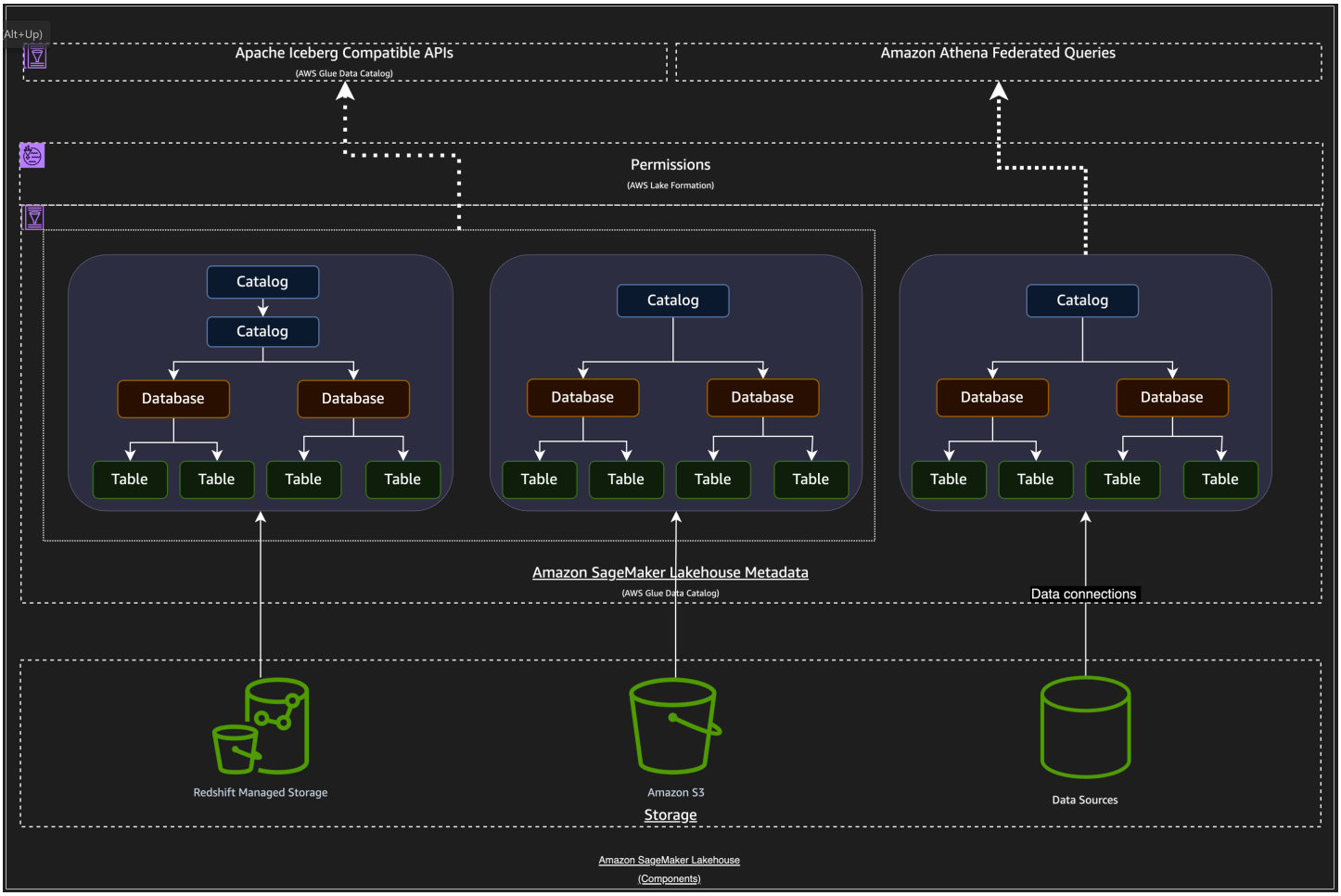
How the lakehouse architecture of Amazon SageMaker works

The lakehouse architecture is accessible from Amazon SageMaker Unified Studio. It organizes data from various sources into logical containers called catalogs. Each catalog represents data from existing sources like Amazon Redshift data warehouses, Amazon S3 data lakes, databases, or enterprise applications. You can also create new catalogs in the lakehouse to store data in S3 or Redshift Managed Storage (RMS).

You can access the data as Apache Iceberg tables and query it using any Iceberg-compatible engine, such as Apache Spark, Amazon Athena, or Amazon EMR. Additionally, these catalogs are mounted as databases in Amazon Redshift, so you can connect and analyze your lakehouse data using SQL tools.

The lakehouse architecture is built on AWS Glue Data Catalog and AWS Lake Formation in your AWS account. With thelakehouse architecture, you can access and query your existing data in Amazon Redshift data warehouses and store new data in RMS from any Apache Iceberg compatible engine.

The following diagram shows how the lakehouse architecture works. Catalogs contain databases, which then contain tables. Types of storage sources for data that goes into catalogs include Redshift Managed Storage, Amazon S3, and data sources that you connect to with data connections.



Data connections in the lakehouse architecture of Amazon SageMaker

The lakehouse architecture provides a unified approach to managing data connections across AWS services and enterprise applications. These connections provide a consistent experience for creating, testing, and exploring data sources, regardless of the underlying data platform.

Capabilities

With the lakehouse architecture connections, you can do the following:

- Create connections to a variety of data sources, including databases and data lakes
- Manage data connections in a single place
- Test the connectivity of your data sources to ensure they are working as expected
- Browse the metadata and preview the data from your connected sources

- Reuse the same connection across different AWS services like AWS Glue, Amazon Athena and Amazon SageMaker AI
- Manage credentials using AWS Secrets Manager
- Authenticate using basic authentication methods such as OAuth2 and IAM

Supported data sources

The lakehouse architecture connections support several popular data sources, including the following:

Supported Data Sources

Data Source	Type
Google BigQuery	Database
Amazon DocumentDB	Database
Amazon DynamoDB	Database
Amazon Redshift	Database
MySQL	Database
PostgreSQL	Database
SQL Server	Database
Snowflake	Database
Oracle	Database
Aurora MySQL	Database
Aurora Postgres	Database
Microsoft Azure SQL	Database

> **Note**
>
> The lakehouse architecture currently supports lowercase table, column, and database names. For optimal experience in the lakehouse architecture, ensure that all database identifiers are in lowercase.

Use the lakehouse architecture connections

After you've created the lakehouse architecture connection, you can use it in various AWS services:

- Amazon SageMaker Unified Studio – Browse metadata, preview sample data, and run SQL queries against the connected data.
- AWS Glue – Use the connection for ETL jobs and crawlers.
- Amazon Athena – Query data directly using Athena's federated query capabilities. For more information, see Register federated catalogs in Amazon Athena.
- Amazon SageMaker AI – Access data for building machine learning models.

Understanding created AWS resources

When you create a connection in Amazon SageMaker Unified Studio, several resources are created in your AWS account(s) behind the scenes. These resources can include:

- AWS Glue connection - A connection object is created in the AWS Glue crawler. This stores the core connection information and is used by various AWS services.
- Athena data catalog - For connections that will be used with Athena , an Athena data catalog is created. This allows Athena to query the external data source.
- AWS Glue data catalog entries - Databases, tables, and schemas from your external data source are registered in the Data Catalog. This enables AWS services to understand the structure of your external data.
- Lambda (for Athena Federated Query) - For some data sources, a Lambda function is created to facilitate federated queries. This function acts as a bridge between Athena and the external data source.

To view these resources, access the respective AWS service consoles (AWS Glue, Athena, IAM, etc.) in the AWS account associated with your Amazon SageMaker Unified Studio project.

In these consoles, look for resources with names that include your Amazon SageMaker Unified Studio project ID or connection name.

For more information about how to create a data connection and explore a connected data source, see ???.

Apache Iceberg support in the lakehouse architecture of Amazon SageMaker

The lakehouse architecture provides comprehensive support for Apache Iceberg, enabling organizations to unify data across Amazon S3 data lakes and Amazon Redshift data warehouses while building powerful analytics and AI/ML applications on a unified data layer.

With the lakehouse architecture, you gain the flexibility to access and query your data in-place using all Apache Iceberg compatible tools and engines, including open-source Apache Spark. This integration leverages the AWS Glue Iceberg REST Catalog, which provides a standardized REST API interface for managing Iceberg table metadata and enables seamless connectivity with third-party engines. For more information, see how to use AWS Glue Iceberg Rest Catalog for accessing Iceberg tables in Amazon S3.

Through fine-grained permissions enforced across all analytics and ML tools, the lakehouse architecture ensures secure data access while supporting advanced Iceberg features like ACID transactions, schema evolution, time travel queries, and efficient row-level operations—all essential capabilities for modern data-driven organizations seeking to process and analyze vast amounts of information efficiently.

The lakehouse architecture also supports multiple table optimization options with Glue Catalog to enhance the management and performance of Apache Iceberg tables that the AWS analytical engines and ETL jobs uses. These optimizers provide efficient storage utilization, improved query performance, and effective data management. For more information, see Optimizing Iceberg tables.

With the lakehouse architecture, you can calculate and update number of distinct values (NDVs) for each column in Iceberg tables with Glue Catalog. These statistics can facilitate better query optimization, data management, and performance efficiency for data engineers and scientists working with large-scale datasets. For more information, see Optimizing query performance for Iceberg tables.

Getting started with the lakehouse architecture of Amazon SageMaker

This guide helps you accomplish common tasks like finding relevant datasets, running SQL queries against your data warehouse and data lake simultaneously, collaborating with team members through data publishing, and maintaining data governance standards. Your administrator will provide the necessary access permissions and project roles to get started.

Topics

Prerequisites

- Your administrator must grant you access to the lakehouse architecture.

  If you don't have access to it, contact your administrator. For more information, see https://docs.aws.amazon.com/sagemaker-unified-studio/latest/userguide/getting-started-access-the-portal.html.

- You must have a Amazon SageMaker Unified Studio project and with the proper project membership role.

  If you don't have proper access to a project, contact your administrator. To view your project membership role, choose **Actions** on the top right corner of the project overview page, then choose **Manage members**. You will see your membership role in the **Role** column.

Create a project

You can create a project from a project profile, which defines a template for projects in your domain. To use lakehouse architecture, your project must be created using either Data analytics and AI-ML model development or SQL analytics project profile. For more information about creating a project, see Create a project from lakehouse architecture User Guide.

When using lakehouse architecture, you can create the following resources in the lakehouse:

1. **Databases in AWS Glue Data Catalog**

   lakehouse architecture is implemented on AWS Glue and AWS Lake Formation in your AWS account.

2. **A catalog to store data in Redshift Managed Storage (RMS) format**

   You will create a catalog in RMS format. To view the catalog, navigate to the AWS Lake Formation console at https://console.aws.amazon.com/lakeformation/, you should be able to see the catalog from the **Catalogs** list.

3. **Provisioning permissions**

   You will create an IAM role when you create a project. Each project has a dedicated IAM role. This IAM role has permission to the resources that are created from this project. The Amazon Resource Name (ARN) of this IAM role is visible from **Project details** section of the **Project overview** page.

Browse data

You can browse data in lakehouse architecture by completing the following steps.

To browse data

1. Choose a project to view the data.
2. On project page, from the left navigation, choose **Data**. This opens the **Data** explorer in the middle of the page.

   The **Data** explorer includes: **Lakehouse**, **Redshift**, and **S3**.

3. Expand **Lakehouse** to view catalogs, databases, tables.

Upload data

You can upload data in CSV or JSON format to a catalog. To upload data, follow the instructions in ???.

After uploading data is complete, you will see the table listed within the database under **AwsDataCatalog**.

Query data

You can query data using supported query editor.

To query data

1. On **Lakehouse**, choose **AwsDataCatalog** on top. Expand the catalog to view the list of databases. Choose a database.
2. From a selected database, choose a table. Then choose the three dot menu to the right of the table to view supported tools for data query.
3. Choose **Query with Athena**. This opens the **Data explorer** page where you can run SQL queries. You might find information in SQL reference for Athena helpful.
4. Choose **Query with Amazon Redshift**. This opens the **Data explorer** page where you can run SQL queries. You might find information in Querying a database using the query editor v2 helpful.

To subscribe an asset, see Request subscription to assets in Amazon SageMaker Unified Studio.

To publish data to the catalog from the lakehouse inventory, see ???.

Adding data sources in lakehouse architecture

lakehouse architecture supports several data sources. If you are new to data connections in lakehouse architecture, see ???.

In this topic:

- Uploading data
- Creating a catalog
- Adding existing databases and catalogs using AWS Lake Formation permissions
- Amazon S3 tables integration

Creating connections in lakehouse architecture

Amazon SageMaker Unified Studio provides an interface for managing and utilizing data connections across various AWS services and external data sources. With Amazon SageMaker Unified Studio, you create, configure, and manage connections to databases, data warehouses, and applications all from a single platform. Amazon SageMaker Unified Studio allows you to explore your connected data sources, preview sample data, and seamlessly use these connections in SQL queries and Spark notebooks without having to switch between different interfaces or manage complex connection details manually.

Access the data explorer in a project

1. Open your web browser and navigate to Amazon SageMaker Unified Studio.
2. Enter your corporate credentials (usually integrated with Amazon IAM Identity Center).
3. After successful authentication, you'll be directed to the Amazon SageMaker Unified Studio home page. On the home page, you'll see a list of projects you have access to. Select the project you want to work with by clicking on its name.
4. From the dropdown menu, select the **Data** or **Data Management** option. This will open the Data section of the project overview page. In this data explorer, you can see a tree-like structure representing your data sources.

Create a new connection to add data sources

To add a new data source

1. In the data explorer, select the **+** button. Click this button to start adding a new data source.
2. In the modal, select **Add connection**. You'll be presented with a gallery of connector options. Select the connector you need. For supported data sources, see .

> **Note**
>
> lakehouse architecture currently supports lowercase table, column, and database names. For optimal experience in lakehouse architecture, ensure that all database identifiers are in lowercase.

3. You must configure your connector details. For example, if you choose to use a DynamoDB connection (preview), fill in the required fields, which can include:

   - Name: A unique identifier for this connection in Amazon SageMaker Unified Studio.

   - Description (optional): A description of the connection.

> **Note**
>
> Each supported data source can have different parameters for the connection. Contact your administrator if you need them.

To see your DynamoDB tables displayed in lakehouse architecture after you add the connection, your administrator must grant you access through resource policies in the Amazon DynamoDB console.

To grant access to a DynamoDB table, your administrator can complete the following steps.

1. Sign in to the AWS Management Console and open the Amazon DynamoDB console at https://console.aws.amazon.com/dynamodb/.
2. On the left navigation of the DynamoDB console, choose **Tables**.
3. From the **Tables** page, choose the table to add access to.
4. On the details page of the selected table, choose **Permission**.
5. On the **Resource-based policy for table** section, update the policy with the project role ARN in `Condition`.

> **Note**
>
> You can find the project ARN on the Page details page in the lakehouse architecture.

The following is an example policy. It allows access of the IAM role named `datazone_user_role_projectid` to perform the allowed actions (`Query`, `Scan`, `DescribeTable`, `PartiQLSelect`) on the specified DynamoDB table. Administrators should choose to allow or deny the set of actions.

```
{
    "Sid": "Statement1",
    "Effect": "Allow",
    "Principal": "*",
    "Action": [
        "dynamodb:Query",
        "dynamodb:Scan",
        "dynamodb:DescribeTable",
        "dynamodb:PartiQLSelect"
     ],
    "Resource": "arn:aws:dynamodb:region:account:table/table_name",
    "Condition": {
        "ArnEquals": {
        "aws:PrincipalArn": "arn:aws:iam::region:role/datazone_user_role_projectid"
        }
    }
}
```

Explore a connected data source

After you have connected your data source, you can explore the data source in the data explorer.

1. After your connection is created, return to the data explorer.
2. You should now see your new connection listed in **Lakehouse**.
3. Expand the new connection to view available databases.
4. Expand a database to explore its schema.
5. You can select a table name to view more details about that table, such as Schema details and a list of tables. You can then examine the tables themselves by selecting a table.

6. You will be able to see tabs for **Columns** and **Sample data**. In the **Columns** view, you can view a list of columns in the table, as well as the data types for each column. In the **Sample data** view, you can see the rows of data from the table and use built-in sorting and filtering options to explore the data.

Authentication and tagging for creating connections

You administrator must create credentials and configure the secret tags for you before you create a connection.

Credentials

When creating a connection, if you choose a data source that requires the credentials for **Authentication**, contact your administrator because they must create and provide these credentials. There are two types of the credentials:

- User name and password
- AWS Secrets Manager

Secret tags

- To ensure the secret can only be used for a particular project, your administrator must tag with the `AmazonDataZoneProject` tag key and the value will be `projectId`.
- To use the secret across multiple projects, your administrator must tag the secret with `for-use-with-all-datazone-projects = true`.

Uploading data

You can upload data to the lakehouse architecture.

To upload data

1. On the **Data** section in the middle of the project page, choose **+** on the top. This opens **Add data source** on the right.
2. On **Add data source**, choose **Upload data**.
3. Choose **Click to upload** or drag and drop a CSV or JSON file. Complete the information in the form.

4. Choose **Upload data**.

Creating a catalog

You can create a catalog for your Redshift Managed Storage (RMS) objects.

To create a catalog

1. On the **Data** explorer in the middle of the project page, choose **+** on the top. This opens **Add data source** on the right.
2. On **Add data source**, choose **Create catalog**. Enter a name for your catalog.
3. Choose **Create**.

Adding existing databases and catalogs using AWS Lake Formation permissions

You can add existing databases and catalogs to the lakehouse architecture.

To add existing databases and catalogs using AWS Lake Formation permissions

1. Sign in to the lakehouse architecture by using the link your administrator gave you. If you don't have access to it, contact your administrator.
2. Choose a project to open the project page.
3. On the left navigation, choose **Project overview**. On **Project details**, copy the project role ARN.
4. Open the AWS Lake Formation console at https://console.aws.amazon.com/lakeformation/.
5. On the left navigation, from **Data catalog**, choose **Catalogs**.
6. On the **Catalogs** list view, choose a catalog you want to add to lakehouse architecture. From **Actions** on the right, choose **Grant**.
7. On the **Grant data lake permissions** page, choose **IAM users and roles** from **Principals**. Paste the IAM role you copied in the step 3.
8. On **Catalog permissions**, choose **Super user**. Choose **Grant**.

After you complete all the steps successfully, go back to the project page in the lakehouse architecture. You should see the Lake Formation catalog added to your lakehouse.

Amazon S3 tables integration

The lakehouse architecture unifies all your data across Amazon S3 data lakes, Amazon Redshift data warehouses, and third-party data sources without having to copy data. Amazon S3 Tables delivers the first cloud object store with built-in Apache Iceberg support. The lakehouse architecture integrates with Amazon S3 Tables so you can access S3 Tables from AWS analytics services, such as Amazon Redshift, Athena, Amazon EMR, AWS Glue, or Apache Iceberg-compatible engines (Apache Spark or PyIceberg).

The lakehouse architecture integration with Amazon S3 Tables helps you secure analytic workflows by joining data from Amazon S3 Tables with sources, such as Amazon Redshift data warehouses, third-party, and federated data sources (Amazon DynamoDB or PostgreSQL). The lakehouse architecture also enables centralized management of fine-grained data access permissions for S3 Tables and other data, and consistently applies them across all engines. To get started, complete the steps in the following sections.

Prerequisites - complete all the steps in the Getting started with the lakehouse architecture of Amazon SageMaker.

Enable Amazon S3 integration

1. Navigate to the Amazon S3 console. In the left navigation pane, choose **Table buckets**.
2. Choose **Create table bucket**.
3. On the **Create table bucket** page, enter a **Table bucket name** and select **Enable integration**.
4. Choose **Create table bucket**.
5. You will see confirmation when Amazon S3 completes integration of your table buckets with the lakehouse architecture.

Onboard S3 Tables in the lakehouse architecture

To provide access to S3 tables, complete the following steps:

1. Navigate to the AWS Lake Formation console.
2. In the left navigation pane, choose **Catalogs** and choose **S3tablescatalog**.
3. From **S3tablescatalog**, under **Objects**, choose the name of your newly created **table bucket**.
4. From the **Actions** menu, select **Grant**.

5. In the **Grant permissions**, under IAM users and roles, select your Amazon SageMaker Unified Studio Project role. To grant full access, under **Catalog Permissions > Grant**, select **Super user**.

Create S3 Table and add data in the lakehouse architecture

1. Navigate to Amazon SageMaker Unified Studio, and select the project.
2. From the **Build** menu, select **Query Editor**, and ensure you have **Athena** selected in **Connections**.
3. Create a database using SQL.

```
CREATE DATABASE "s3tablescatalog/<Your Bucket Name>".<YourDBName>;
```

4. Create an S3 table using SQL.

```
CREATE TABLE "s3tablescatalog/<Your Bucket Name>".<YourDBName>.<YourTableName>
( c_salutation string,
  c_login string,
  c_first_name string,
  c_last_name string,
  c_email_address string)
  TBLPROPERTIES (
  'table_type'='ICEBERG'  );
```

5. Add data using SQL.

```
INSERT INTO "s3tablescatalog/<Your Bucket Name>".<YourDBName>.<YourTableName>
 VALUES('Dr.','1381546','Joyce','Deaton','Joyce.Deaton@qhtrwert.edu');
```

You can now use the following integrated analytics services:

- Amazon Athena - create databases, tables, query and add data in S3 Tables.
- Amazon Redshift - query data from S3 Tables.
- Amazon EMR - create table, namespace, query and add data in S3 Tables.

- AWS Glue - create table, namespace, query and add data in S3 Tables.
- AWS Lake Formation - grant fine-grained permissions for S3 table catalogs, databases, tables, columns, and cells.

> **Note**
>
> Access to S3 Tables with the lakehouse architecture is available in the AWS Regions where S3 Tables are available.

Publishing data in lakehouse architecture

After you have added data in the lakehouse architecture, you can publish the data to share it with other users in the lakehouse architecture. Data that is published is viewable as an asset in the project catalog and the Amazon SageMaker Catalog, and other users can create subscription requests in the Amazon SageMaker Catalog to include that data in their projects.

To publish data in the lakehouse architecture, complete the following steps:

1. Navigate to lakehouse architecture using the URL from your admin and log in using your SSO or AWS credentials.
2. Navigate to the project that contains the data that you want to publish in the lakehouse architecture. To do this, use the center menu at the top of the landing page and choose **Browse all projects**, then choose the name of the project that you want to navigate to.
3. In the center menu, choose **Data**. This takes you to the Data page.
4. Do either of the following:

    - If you want to publish a regular AWS Glue table, expand the catalog in the data navigation to view the list of databases in lakehouse architecture, then choose a database that contains the asset that you want to publish. Choose this table from the selected database and then proceed to the rest of the steps in this procedure to publish this table to the catalog.
    - If you want to publish an Amazon S3 table to the catalog, you must first complete the following steps to create a data source for the S3 Tables catalog and schedule its run job. Then you can proceed to the rest of the steps in this procedure to publish the S3 table to the catalog.
        - Navigate to **Data sources** and then choose **Create data source**.

- On the **Step 1: Define source** page, specify the name for this data source, then under **Data source type** - choose **AWS Glue (Lakehouse)**, under **Data Selection** - choose **Enter the catalog name** and then speciy the name of your S3 tables catalog (s3tablescatalog/<catalog name>, then choose your database from that catalog (use the drop down menu), and then choose **Next**.
- On the **Step 2: Add details** page, leave all the default settings and choose **Next**.
- On the **Step 3: Set up schedule** page, choose a run preference and then choose **Next**.
- On the **Step 4: review** page, review your selections and then choose **Create**.

  Once the data source for the S3 tables catalog is created and run, you can proceed with the rest of the steps below to locate your S3 table and publish it to the catalog.

5. Expand the **Actions** menu, then choose **Publish to catalog**.
6. Confirm the action in the pop-up window by choosing **Publish to catalog**.

   The lakehouse architecture then fetches metadata for the asset. After a few minutes, the metadata is fetched and a success message appears.
7. (Optional) Choose **View details** to view the asset in the project catalog.

When it is successfully published you can view it in the **Assets** section of the project catalog and users in other projects can subscribe to it from the Amazon SageMaker Catalog.

You can use the project catalog to re-publish the data if you make changes, or to unpublish the data from Amazon SageMaker Catalog. For more information, see Data inventory and publishing.

Document history for the lakehouse architecture of Amazon SageMaker User Guide

The following table describes the documentation releases for the lakehouse architecture.

Change	Description	Date
Added new sections	Added the getting started and sections.	August 26, 2025
Initial release	Initial release of the lakehouse architecture of Amazon SageMaker User Guide	July 31, 2025