AWS Decision Guide

# Choosing an AWS analytics service

# Choosing an AWS analytics service: AWS Decision Guide

# Table of Contents

**Decision guide ..... 1**
Introduction ..... 1
Understand ..... 2
Consider ..... 6
Choose ..... 14
Use ..... 19
Explore ..... 30
**Document history ..... 31**

# Choosing an AWS analytics service

**Taking the first step**

| | |
|---|---|
| **Purpose** | Help determine which AWS analytics services are the best fit for your organization. |
| **Last updated** | November 17, 2023 |
| **Covered services** | • Amazon AppFlow<br>• Amazon Athena<br>• AWS Data Exchange<br>• Amazon DataZone<br>• Amazon EMR<br>• AWS Glue<br>• Amazon Managed Service for Apache Flink<br>• AWS Lake Formation<br>• Amazon Managed Streaming for Apache Kafka (Amazon MSK)<br>• OpenSearch<br>• Amazon QuickSight<br>• Amazon Redshift<br>• Amazon S3<br>• Amazon SageMaker |

## Introduction

Data needs to be securely accessed and analyzed by applications and people. Data volumes are coming from new and diverse sources, and increasing at an unprecedented rate. Organizations need to extract data value, but struggle to capture, store, and analyze all the data generated by today's modern businesses.

Meeting these challenges means building a modern data architecture that breaks down all of your data silos for analytics and insights—including third-party data—and puts it in the hands of everyone in the organization, with end-to-end governance. It is also increasingly important to connect your analytics and machine learning (ML) systems to enable predictive analytics.

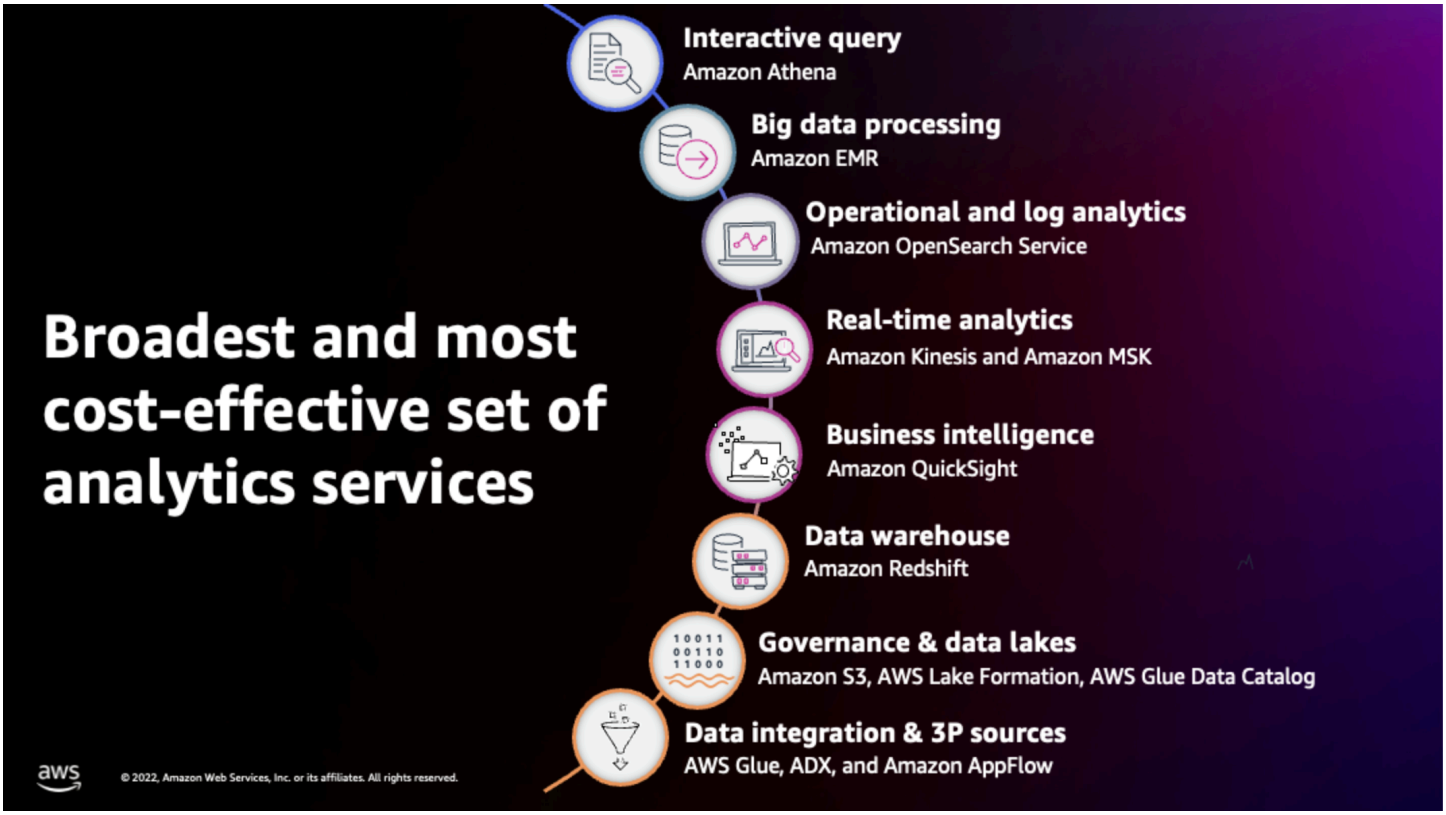
This decision guide helps you ask the right questions to build your modern data architecture on AWS services. It explains how to break down your data silos (by connecting your data lake and data warehouses), your system silos (by connecting ML and analytics), and your people silos (by putting data in the hands of everyone in your organization).

*This six-minute excerpt is from a one-hour presentation by G2 Krishnamoorthy, VP of AWS Analytics at re:Invent 2022. It provides an overview of AWS analytics services. The full presentation covers the current state of analytics on AWS as well as the latest service innovations around data, and highlights customer successes with AWS analytics.*

# Understand

A modern data strategy is enabled by a set of technology building blocks that help you manage, access, analyze, and act on data. It also gives you multiple options to connect to data sources. A modern data strategy should empower your teams to:

- Run analytics or ML using your preferred tools or techniques
- Manage who has access to data with the proper security and data governance controls
- Break down data silos to give you the best of both data lakes and purpose-built data stores
- Store any amount of data, at low-cost, and in open, standards-based data formats. The AWS modern data architecture connects your lake, warehouse, and other purpose-built services into a coherent whole.

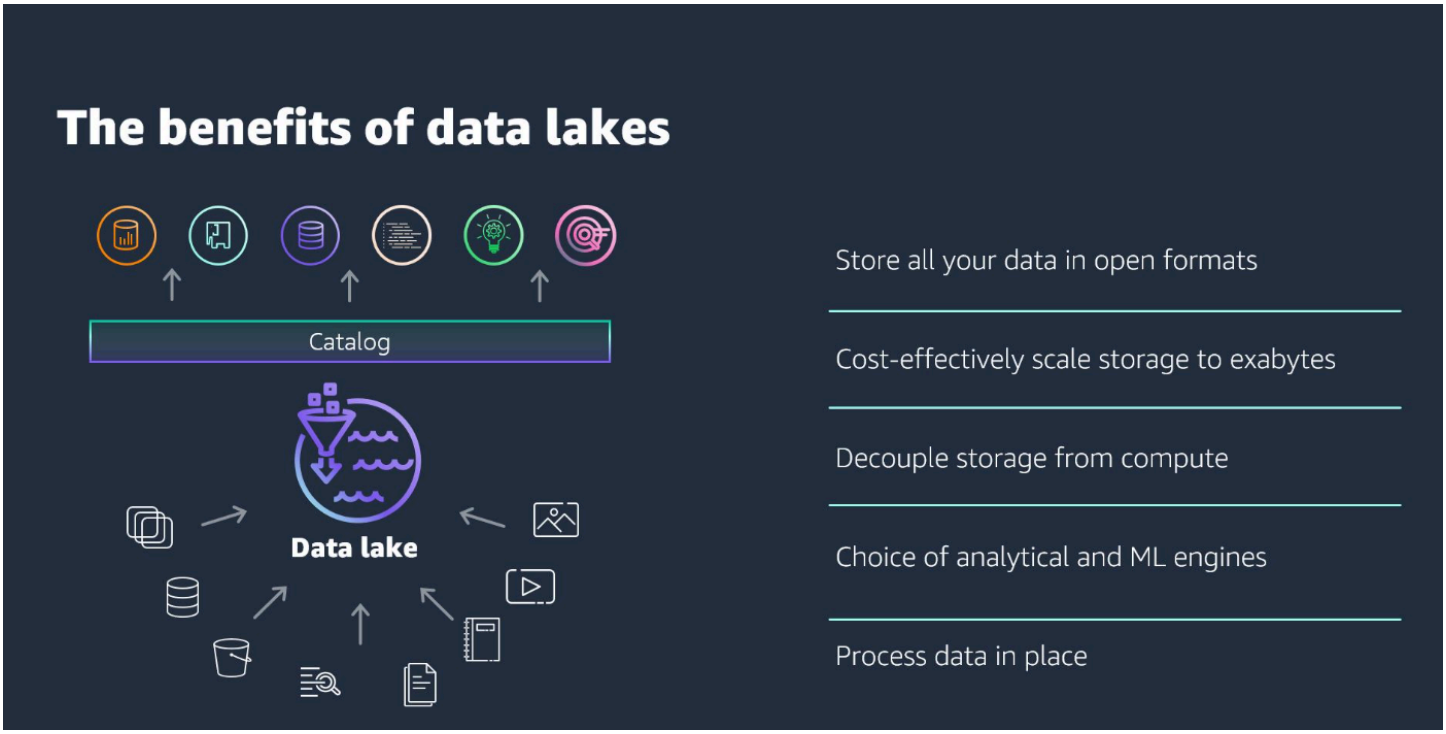



**Implementing a modern data strategy on AWS is based on the following five pillars:**


## Scalable data lakes

To make decisions quickly, you will want to store any amount of data in open formats and be able to break down disconnected data silos. You might also have a need to empower people in your organization to run analytics or ML (using your preferred tools or techniques for doing so), as well as manage who can access specific pieces of data with the proper security and data governance controls.

A modern data architecture starts with the data lake. A data lake lets you store all of your data (relational, non-relational, structured, and unstructured) cost effectively. With AWS, you can move any amount of data from various silos into an Amazon S3 data lake. Amazon S3 then stores data using a standard-based open format.




**Purpose-built for performance and cost**

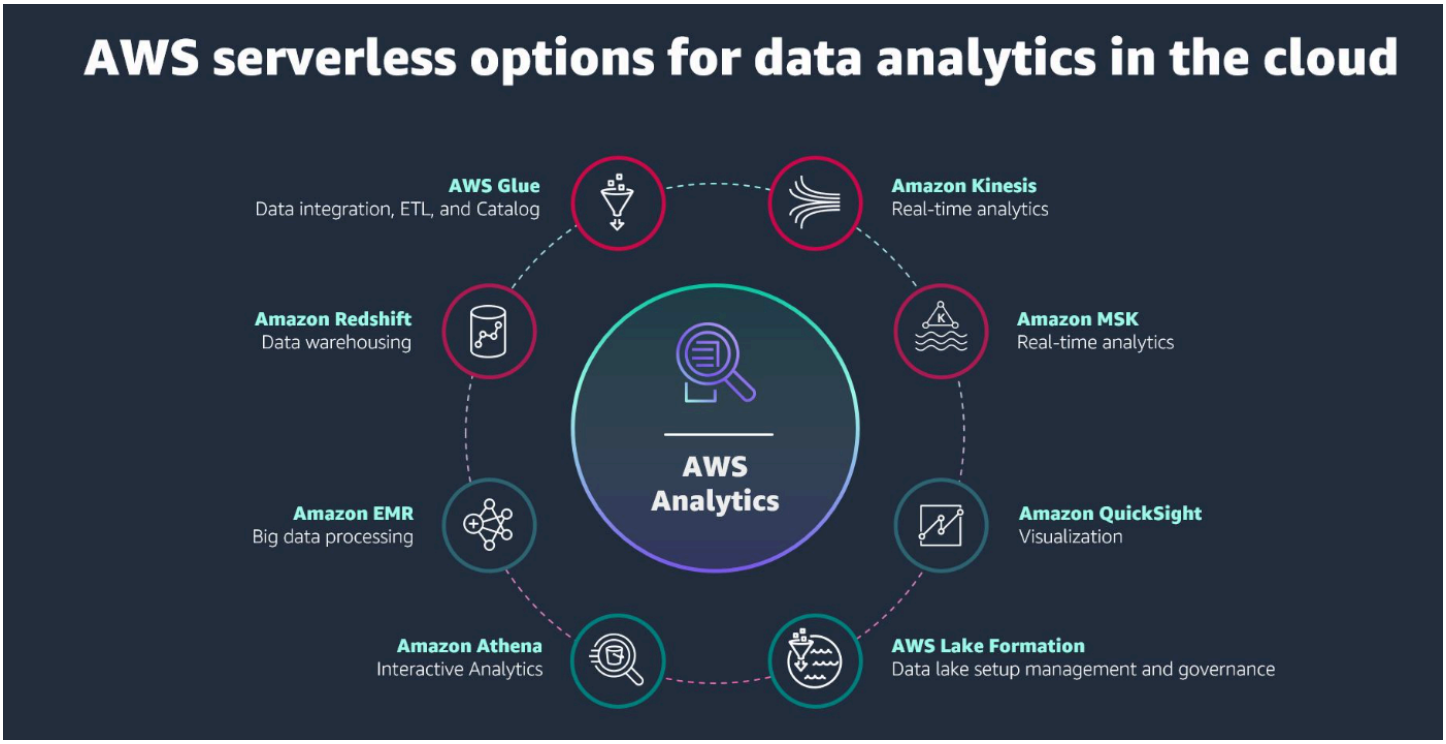
On-premises data pipelines are often retrofitted to the tools you are currently using, providing a sub-optimal experience. AWS provides a broad and deep set of purpose-built data services allowing you to choose the right tool for the right job so you don't have to compromise on functionality, performance, scale, or cost.

**Serverless and easy to use**

For many types of analytics needs, AWS provides serverless options designed to enable your to focus on your application, without having to touch any infrastructure.

The process of getting raw data into a state that can be used to derive business insights, and performed by the extract, transform, and load (ETL) stage of the data pipeline, can be challenging. AWS is moving towards a Zero-ETL approach (one that eliminates the need for traditional ETL processes). This approach will help you analyze data where it sits, without the need to use ETL. Features within AWS services that support this approach include:

- Amazon Zero-ETL Aurora to Amazon Redshift
- Amazon Redshift Streaming Ingestion directly from Kinesis and Amazon MSK to Redshift
- Federated Query in Amazon Redshift and Amazon Athena




**Unified data access, security, and governance**

Once you have a centralized data lake and collection of purpose-built analytics services, you then need the ability to access that data wherever it lives, then secure it and have governance policies to comply with relevant regulations and security best practices.

Governance starts with AWS Lake Formation. This service allows you to access your data wherever it lives, whether it's in a database, data warehouse, purpose-built data store, or a data lake, and then keep your data secure no matter where you store it.

For data governance, AWS automatically discovers, tags, catalogs, and keeps your data in sync and you can centrally define and manage security, governance, and auditing policies to satisfy regulations specific to your industry and geography.

**Built-in machine learning**

AWS offers built-in ML integration as part of our purpose-built analytics services. You can build, train, and deploy ML models using familiar SQL commands, without any prior ML experience.


It is not uncommon to use different types of data stores (relational, non-relational, data warehouses, and analytics services) for different use cases. AWS provides a range of integrations to give you options for training models on your data—or adding inference results right from your data store—without having to export and process your data.

## Consider

There are many reasons for building an analytics pipeline on AWS. You may need to support a greenfield or pilot project as a first step in your cloud migration journey. Alternatively, you may be migrating an existing workload with as little disruption as possible. Whatever your goal, the following considerations may be useful in making your choice.

### Assess data sources and data types

Analyze available data sources and data types to gain a comprehensive understanding of data diversity, frequency, and quality. Understand any potential challenges in processing and analyzing the data. This analysis is crucial because:

- Data sources are diverse and come from various systems, applications, devices, and external platforms.
- Data sources have unique structure, format, and frequency of data updates. Analyzing these sources helps in identifying suitable data collection methods and technologies.
- Analyzing data types, such as structured, semi-structured, and unstructured data determines the appropriate data processing and storage approaches.
- Analyzing data sources and types of facilitates data quality assessment, helps you anticipate potential data quality issues—missing values, inconsistencies, or inaccuracies.

### Data processing requirements

Determine data processing requirements for how data is ingested, transformed, cleansed, and prepared for analysis. Key considerations include:

- **Data transformation:** Determine the specific transformations needed to make the raw data suitable for analysis. This involves tasks like data aggregation, normalization, filtering, and enrichment.
- **Data cleansing:** Assess data quality and define processes to handle missing, inaccurate, or inconsistent data. Implement data cleansing techniques to ensure high-quality data for reliable insights.
- **Processing frequency:** Determine whether real-time, near real-time, or batch processing is required based on the analytical needs. Real-time processing enables immediate insights, while batch processing may be sufficient for periodic analyses.
- **Scalability and throughput:** Evaluate the scalability requirements for handling data volumes, processing speed, and the number of concurrent data requests. Ensure that the chosen processing approach can accommodate future growth.
- **Latency:** Consider the acceptable latency for data processing and the time it takes from data ingestion to analysis results. This is particularly important for real-time or time-sensitive analytics.

### Storage requirements

Determine storage needs by determining how and where data is stored throughout the analytics pipeline. Important considerations include:

- **Data volume:** Assess the amount of data being generated and collected, and estimate future data growth to plan for sufficient storage capacity.
- **Data retention:** Define the duration for which data should be retained for historical analysis or compliance purposes. Determine the appropriate data retention policies.
- **Data access patterns:** Understand how data will be accessed and queried to choose the most suitable storage solution. Consider read and write operations, data access frequency, and data locality.
- **Data security:** Prioritize data security by evaluating encryption options, access controls, and data protection mechanisms to safeguard sensitive information.
- **Cost optimization:** Optimize storage costs by selecting the most cost-effective storage solutions based on data access patterns and usage.
- **Integration with analytics services:** Ensure seamless integration between the chosen storage solution and the data processing and analytics tools in the pipeline.

### Types of data

When deciding on analytics services for the collection and ingestion of data, consider various types of data that are relevant to your organization's needs and objectives. Common types of data that you might need to consider includes:

- **Transactional data:** Includes information about individual interactions or transactions, such as customer purchases, financial transactions, online orders, and user activity logs.
- **File-based data:** Refers to structured or unstructured data that is stored in files, such as log files, spreadsheets, documents, images, audio files, and video files. Analytics services should support the ingestion of different file formats/
- **Event data:** Captures significant occurrences or incidents, such as user actions, system events, machine events, or business events. Events can include any data that is arriving in high velocity that is captured for on-stream or down-stream processing.

### Operational considerations

Operational responsibility is shared between you, and AWS, with the division of responsibility varying across different levels of modernization. You have the option of self-managing your analytics infrastructure on AWS or leveraging the numerous serverless analytics services to lesson your infrastructure management burden.

Self-managed options grant users greater control over the infrastructure and configurations, but they require more operational effort.

Serverless options abstract away much of the operational burden, providing automatic scalability, high availability, and robust security features, allowing users to focus more on building analytical solutions and driving insights rather than managing infrastructure and operational tasks. Consider these benefits of serverless analytics solutions:

- **Infrastructure abstraction:** Serverless services abstract infrastructure management, relieving users from provisioning, scaling, and maintenance tasks. AWS handles these operational aspects, reducing management overhead.
- **Auto-Scaling and performance:** Serverless services automatically scale resources based on workload demands, ensuring optimal performance without manual intervention.
- **High availability and disaster recovery:** AWS provides high availability for serverless services. AWS manages data redundancy, replication, and disaster recovery to enhance data availability and reliability.
- **Security and compliance:** AWS manages security measures, data encryption, and compliance for serverless services, adhering to industry standards and best practices.
- **Monitoring and logging:** AWS offers built-in monitoring, logging, and alerting capabilities for serverless services. Users can access detailed metrics and logs through Amazon CloudWatch.

### Type of workload

When building a modern analytics pipeline, deciding on the types of workload to support is crucial to meet different analytical needs effectively. Key decision points to consider for each type of workload includes:

#### Batch workload

- **Data volume and frequency:** Batch processing is suitable for large volumes of data with periodic updates.
- **Data latency:** Batch processing might introduce some delay in delivering insights compared to real-time processing.

**Interactive analysis**

- **Data query complexity:** Interactive analysis requires low-latency responses for quick feedback.
- **Data visualization:** Evaluate the need for interactive data visualization tools to enable business users to explore data visually.

**Streaming workloads**

- **Data velocity and volume:** Streaming workloads require real-time processing to handle high-velocity data.
- **Data windowing:** Define data windowing and time-based aggregations for streaming data to extract relevant insights.

Type of analysis needed

Clearly define the business objectives and the insights you aim to derive from the analytics. Different types of analytics serve different purposes. For example:

- Descriptive analytics is ideal for gaining a historical overview
- Diagnostic analytics helps understand the reasons behind past events
- Predictive analytics forecasts future outcomes
- Prescriptive analytics provides recommendations for optimal actions

Match your business goals with the relevant types of analytics. Here are some key decision criteria to help you choose the right types of analytics:

- **Data availability and quality:** Descriptive and diagnostic analytics rely on historical data, while predictive and prescriptive analytics require sufficient historical data and high-quality data to build accurate models.

- **Data volume and complexity:** Predictive and prescriptive analytics require substantial data processing and computational resources. Ensure that your infrastructure and tools can handle the data volume and complexity.
- **Decision complexity:** If decisions involve multiple variables, constraints, and objectives, prescriptive analytics may be more suitable to guide optimal actions.
- **Risk tolerance:** Prescriptive analytics may provide recommendations, but come with associated uncertainties. Ensure that decision-makers understand the risks associated with the analytics outputs.

### Evaluate scalability and performance

Assess the scalability and performance needs of the architecture. The design must handle increasing data volumes, user demands, and analytical workloads. Key decision factors to consider includes:

- **Data volume and growth:** Assess the current data volume and anticipate future growth.
- **Data velocity and real-time requirements:** Determine if the data needs to be processed and analyzed in real-time or near real-time.
- **Data processing complexity:** Analyze the complexity of your data processing and analysis tasks. For computationally intensive tasks, services such as Amazon EMR provide a scalable and managed environment for big data processing.
- **Concurrency and user load:** Consider the number of concurrent users and the level of user load on the system.
- **Auto-scaling capabilities:** Consider services that offer auto-scaling capabilities, allowing resources to automatically scale up or down based on demand. This ensures efficient resource utilization and cost optimization.
- **Geographic distribution:** Consider services with global replication and low-latency data access if your data architecture needs to be distributed across multiple regions or locations.
- **Cost-performance trade-off:** Balance the performance needs with cost considerations. Services with high performance may come at a higher cost.
- **Service-level agreements (SLAs):** Check the SLAs provided by AWS services to ensure they meet your scalability and performance expectations.

### Data governance

Data governance is the set of processes, policies, and controls you need to implement to ensure effective management, quality, security, and compliance of your data assets. Key decision points to consider includes:

- **Data retention policies:** Define data retention policies based on regulatory requirements and business needs and establish processes for secure data disposal when it is no longer needed.
- **Audit trail and logging:** Decide on the logging and auditing mechanisms to monitor data access and usage. Implement comprehensive audit trails to track data changes, access attempts, and user activities for compliance and security monitoring.
- **Compliance requirements:** Understand the industry-specific and geographic data compliance regulations that apply to your organization. Ensure that the data architecture aligns with these regulations and guidelines.
- **Data classification:** Classify data based on its sensitivity and define appropriate security controls for each data class.
- **Disaster recovery and business continuity:** Plan for disaster recovery and business continuity to ensure data availability and resilience in case of unexpected events or system failures.
- **Third-party data sharing:** If sharing data with third-party entities, implement secure data sharing protocols and agreements to protect data confidentiality and prevent data misuse.

### Security

The security of data in the analytics pipeline involves protecting data at every stage of the pipeline to ensure its confidentiality, integrity, and availability. Key decision points to consider includes:

- **Access control and authorization:** Implement robust authentication and authorization protocols to ensure that only authorized users can access specific data resources.
- **Data encryption:** Choose appropriate encryption methods for data stored in databases, data lakes, and during data movement between different components of the architecture.
- **Data masking and anonymization:** Consider the need for data masking or anonymization to protect sensitive data, such as PII or sensitive business data, while allowing certain analytical processes to continue.

- **Secure data integration:** Establish secure data integration practices to ensure that data flows securely between different components of the architecture, avoiding data leaks or unauthorized access during data movement.
- **Network isolation:** Consider services that support Amazon VPC Endpoints to avoid exposing resources to the public internet.

### Plan for integration and data flows

Define the integration points and data flows between various components of the analytics pipeline to ensure seamless data flow and interoperability. Key decision points to consider includes:

- **Data source integration:** Identify the data sources from which data will be collected, such as databases, applications, files, or external APIs. Decide on the data ingestion methods (batch, real-time, event-based) to bring data into the pipeline efficiently and with minimal latency.
- **Data transformation:** Determine the transformations required to prepare data for analysis. Decide on the tools and processes to clean, aggregate, normalize, or enrich the data as it moves through the pipeline.
- **Data movement architecture:** Choose the appropriate architecture for data movement between pipeline components. Consider batch processing, stream processing, or a combination of both based on the real-time requirements and data volume.
- **Data replication and sync:** Decide on data replication and synchronization mechanisms to keep data up-to-date across all components. Consider real-time replication solutions or periodic data syncs depending on data freshness requirements.
- **Data quality and validation:** Implement data quality checks and validation steps to ensure the integrity of data as it moves through the pipeline. Decide on the actions to be taken when data fails validation, such as alerting or error handling.
- **Data security and encryption:** Determine how data will be secured during transit and at rest. Decide on the encryption methods to protect sensitive data throughout the pipeline, considering the level of security required based on data sensitivity.
- **Scalability and resilience:** Ensure that the data flow design allows for horizontal scalability and can handle increased data volumes and traffic.

### Architect for cost optimization

Building your analytics pipeline on AWS provides various cost optimization opportunities. To ensure cost efficiency, consider the following strategies:

- **Resource sizing and selection:** Right-size your resources based on actual workload requirements. Choose AWS services and instance types that match the workloads performance needs while avoiding overprovisioning.
- **Auto-scaling:** Implement auto-scaling for services that experience varying workloads. Auto-scaling dynamically adjusts the number of instances based on demand, reducing costs during low-traffic periods.
- **Spot Instances:** Use Amazon EC2 Spot Instances for non-critical and fault-tolerant workloads. Spot Instances can significantly reduce costs compared to on-demand instances.
- **Reserved instances:** Consider purchasing AWS Reserved Instances to achieve significant cost savings over on-demand pricing for stable workloads with predictable usage.
- **Data storage tiering:** Optimize data storage costs by using different storage classes based on data access frequency.
- **Data lifecycle policies:** Establish data lifecycle policies to automatically move or delete data based on its age and usage patterns. This helps manage storage costs and keeps data storage aligned with its value.

## Choose

Now that you know the criteria to evaluate your analytics needs, you are ready to choose which AWS analytics services are right for your organizational needs. The following table categorizes sets of services aligning to what you will need to accomplish with for your business goals, such as conducting advanced analytics, performing data management or predictive analytics, and ML.

| Categories | What is it optimized for? | Services |
| --- | --- | --- |
| **Advanced analytics** | | |
| | **Interactive analytics**<br><br>Optimized for performing real-time data analysis and exploration, which allows | Amazon Athena |

| Categories | What is it optimized for? | Services |
| --- | --- | --- |
| | users to interactively query and visualize data to gain insights and make data-driven decisions quickly. | |
| | **Big data processing**<br>Big data is characterized by its three dimensions, volume, velocity, and variety. Big data processing solutions aim to overcome the challenges posed by the sheer scale and complexity of big data. | Amazon EMR |
| | **Data warehousing**<br>The centralized storage, organization, and retrieval of large volumes of structured and sometimes semi-structured data from various sources within an organization. | Amazon Redshift |
| | **Real-time analytics**<br>The process of analyzing and processing data as it is generated, received, or ingested, without any significant delay. | Amazon Managed Service for Apache Flink |

| Categories | What is it optimized for? | Services |
| --- | --- | --- |
| | **Operational analytics**<br><br>The use of real-time data analysis and insights to optimize and improve ongoing operational processes and activities within an organization. | Amazon OpenSearch Service |
| | **Dashboards and visualizations**<br><br>Dashboards and visualizations provide a visual representation of complex data sets, making it easier for users to grasp patterns, trends, and insights at a glance. They simplify the understanding of data, even for non-technical users, by presenting information in a visually appealing and intuitive manner. | Amazon QuickSight |
| | **Visual data preparation**<br><br>Using visual tools and interfaces to explore, clean, transform, and manipulate data in a visual and intuitive manner. | AWS Glue DataBrew |
| **Data management** | **Use cases** | **Related analytics services** |

| Categories | What is it optimized for? | Services |
| --- | --- | --- |
| | **Real-time data movement** | Amazon MSK |
| | Real-time data movement involves minimal delay in transferring data, typically within seconds or milliseconds after it becomes available. | Amazon Kinesis Data Streams |
| | | Amazon Data Firehose |
| | | Amazon Kinesis Video Streams |
| | | AWS Glue |
| | **Data governance** | Amazon DataZone |
| | A set of processes, policies, and guidelines that ensure the proper managemen t, availability, usability, integrity, and security of data throughout its lifecycle. | AWS Lake Formation |
| | **Object storage for data lakes** | Amazon S3 |
| | A data lake built on AWS uses Amazon S3 as its primary storage platform. Amazon S3 provides an optimal foundatio n for a data lake because of its virtually unlimited scalabili ty and high durability. | AWS Lake Formation |

| Categories | What is it optimized for? | Services |
| --- | --- | --- |
| | **Backup and archive for data lakes**<br><br>Data lakes, powered by Amazon S3, provide organizations with the availability, agility, and flexibility required for modern analytics approaches to gain deeper insights. Protecting sensitive or business-critical information stored in these S3 buckets is a high priority for organizations. | Amazon S3 Glacier<br><br>AWS Backup |
| | **Data catalog**<br><br>A metadata management tool, providing detailed information about the available data, its structure, characteristics, and relationships. | AWS Glue |
| | **Third-party data**<br><br>Third-party data and software-as-a-service (SaaS) data are becoming increasingly important to business operations in the modern data-driven landscape. | AWS Data Exchange<br><br>Amazon AppFlow |
| **Predictive analytics and machine learning** | **Use cases** | **Related analytics services** |

| Categories | What is it optimized for? | Services |
| --- | --- | --- |
| | **Frameworks and interfaces** | AWS Deep Learning AMIs |
| | AWS ML infrastructure supports all of the leading ML frameworks. | |
| | **Platform services** | Amazon SageMaker |
| | Fully managed infrastructure for building, training, and deploying machine learning models. | |
| | **Direct data integrations** | Amazon Athena ML |
| | Create, train, and deploy ML models using familiar SQL commands. | Amazon QuickSight ML Insights |
| | | Amazon Redshift ML |

# Use

You should now have a clear understanding of your business objectives, and the volume and velocity of data you will be ingesting and analyzing to begin building your data pipelines.

To explore how to use and learn more about each of the available services—we have provided a pathway to explore how each of the services work. The following sections provides links to in-depth documentation, hands-on tutorials, and resources to get you started from basic usage to more advanced deep dives.

Amazon AppFlow

**Getting started with Amazon AppFlow**

Learn how to use Amazon Athena to query data and create a table based on sample data stored in Amazon S3, query the table, and check the results of the query.

Read the guide

**Tutorial: Transfer data between applications with Amazon AppFlow**

In this tutorial, you learn to transfer data between applications. Specifically, you transfer data both from Amazon S3 to Salesforce, and from Salesforce to Amazon S3.

Get started with the tutorial

**Hands On Workshop: Amazon AppFlow Workshop**

You will learn about Amazon AppFlow and how to easily transfer data between popular SaaS services and AWS. The workshop is divided into multiple modules, each targeting a specific SaaS application integration.

Get started with the workshop

Amazon Athena

**Getting started with Amazon Athena**

Learn how to use Amazon Athena to query data and create a table based on sample

**Get started with Apache Spark on Amazon Athena**

Use the simplified notebook experience in Amazon Athena console to develop Apache

data stored in Amazon S3, query the table, and check the results of the query.

Get started with the tutorial

Spark applications using Python or Athena notebook APIs.

Get started with the tutorial

**AWS re:Invent 2022 - What's new in Amazon Athena**

Learn how you can bring Athena to your data, applying it to all of your data spanning data lakes, external sources, and more.

Watch the session

**Analyzing data in S3 using Amazon Athena**

Explore how to use Athena on logs from Elastic Load Balancers, generated as text files in a pre-defined format. We show you how to create a table, partition the data in a format used by Athena, convert it to Parquet, and compare query performance.

Read the blog post

AWS Data Exchange

**Getting started as an AWS Data Exchange subscriber**

Understand the complete process of becoming a data product subscriber on AWS Data Exchange using the AWS Data Exchange console.

Explore the guide

**Getting started as an AWS Data Exchange provider**

Understand the complete process of becoming a data product provider on AWS Data Exchange using the AWS Data Exchange console.

Explore the guide

**AWS Data Exchange workshop**

Explore self-service labs that you can use to understand and learn how AWS services can be used in conjunction with third-party data to add insights to your data analytics projects.

Get started with the workshop

Amazon DataZone

**Getting started with Amazon DataZone**

Learn how to create the Amazon DataZone root domain, obtain the data portal URL, walk through the basic Amazon DataZone workflows for data producers and data consumers.

Get started with the tutorial

Amazon EMR

**Getting started with Amazon EMR**

Learn how to launch a sample cluster using Spark, and how to run a simple PySpark script stored in an Amazon S3 bucket.

**Getting started with Amazon EMR on EKS**

We show you how to get started using Amazon EMR on EKS by deploying a Spark application on a virtual cluster.

Get started with the tutorial

Explore the guide

**Get started with EMR Serverless**

Explore how Amazon EMR Serverless provides a serverless runtime environment that simplifies the operation of analytics applications that use the latest open source frameworks.

Get started with the tutorial

**What's new with Amazon EMR**

Learn about the latest Amazon EMR developments, including Amazon EMR Serverless, Amazon EMR Studio, and more.

Watch the session

AWS Glue

**Getting started with AWS Glue DataBrew**

Learn how to create your first DataBrew project. You load a sample dataset, run transformations on that dataset, build a recipe to capture those transformations, and run a job to write the transformed data to Amazon S3.

Get started with the tutorial

**Transform data with AWS Glue DataBrew**

Learn about AWS Glue DataBrew, a visual data preparation tool that makes it easy for data analysts and data scientists to clean and normalize data to prepare it for analytics and machine learning. Learn how to construct an ETL process using AWS Glue DataBrew.

Get started with the lab

**AWS Glue DataBrew immersion day**

Explore how to use AWS Glue DataBrew to clean and normalize data for analytics and machine learning.

Get started with the workshop

**Getting started with the AWS Glue Data Catalog**

Learn how to create your first AWS Glue Data Catalog, which uses an Amazon S3 bucket as your data source.

Get started with the tutorial

**Data catalog and crawlers in AWS Glue**

Discover how you can use the information in the Data Catalog to create and monitor your ETL jobs.

Explore the guide

Amazon Managed Service for Apache Flink

**Getting started with Amazon Managed Service for Apache Flink**

Understand the fundamental concepts of Amazon Managed Service for Apache Flink.

Explore the guide

**Streaming analytics workshop**

Learn how to build an end-to-end streaming architecture to ingest, analyze, and visualize streaming data in near real-time.

Get started with the workshop

**Amazon Managed Service for Apache Flink Workshop**

In this workshop, you will learn how to deploy, operate, and scale a Flink application with Amazon Managed Service for Apache Flink.

Attend the virtual workshop

AWS Lake Formation

**Getting started with AWS Lake Formation**

Learn how to set up AWS Lake Formation.

Get started with the guide

**Tutorial: Creating a data lake from a JDBC source in Lake Formation**

Learn about AWS Glue DataBrew, a visual data preparation tool that makes it easy for data analysts and data scientists to clean and normalize data to prepare it for analytics and machine learning. Learn how to construct an ETL process using AWS Glue DataBrew.

Get started with the tutorial

**AWS Lake Formation workshop**

In this workshop there are a series of labs that you to follow through as four

different personas who can benefit from Lake Formation. The personas are Data Admin, Data Engineer, Data Analyst and Data Scientist.

Get started with the workshop

Amazon Managed Streaming for Apache Kafka (Amazon MSK)

**Getting Started with Amazon Managed Streaming for Apache Kafka (Amazon MSK)**

Learn how to create an Amazon MSK cluster, produce and consume data, and monitor the health of your cluster using metrics.

Get started with the guide

**Amazon MSK Workshop**

Go deep with this hands-on Amazon MSK workshop.

Get started with the workshop

OpenSearch

**Getting started with OpenSearch Service**

Learn how to use Amazon OpenSearc h Service to create and configure a test domain.

Get started with the tutorial

**Visualizing customer support calls with OpenSearch Service and OpenSearch Dashboards**

Discover a full walkthrough of the following situation: a business receives some number of customer support calls and wants to analyze them. What is the subject of each call? How many were positive? How many

were negative? How can managers search or review the the transcripts of these calls?

Get started with the tutorial

**Getting started with OpenSearch Serverless workshop**

Learn how to set up a new Amazon OpenSearch Serverless domain in the AWS console. Explore the different types of search queries available, and design eye-catching visualizations, and learn how you can secure your domain and documents based on assigned user privileges.

Get started with the workshop

**Building a log analytics solution with OpenSearch Service**

Learn how to size an OpenSearch cluster for a log analytics workload.

Read the blog post

Amazon QuickSight

**Getting started with Amazon QuickSight data analysis**

Learn how to create your first analysis. Use sample data to create either a simple or a more advanced analysis. Or you can connect to your own data to create an analysis.

Explore the guide

**Visualizing with QuickSight**

Discover the technical side of business intelligence (BI) and data visualization with AWS. Learn how to embed dashboards into applications and websites, and securely manage access and permissions.

Get started with the course

**QuickSight workshops**

Get a head start on your QuickSight journey with workshops

Get started with the workshops

Amazon Redshift

**Getting started with Amazon Redshift Serverless**

Understand the basic flow of Amazon Redshift Serverless to create serverless resources, connect to Amazon Redshift Serverless, load sample data, and then run queries on the data.

Explore the guide

**Modernize your data warehouse**

Explore how you can use the new capabilities of Amazon Redshift to modernize your data warehouse by gaining access to all your data.

Watch the video

**Deploy a data warehouse on AWS**

Learn how to create and configure an Amazon Redshift data warehouse, load

**Amazon Redshift deep dive workshop**

Explore a series of exercises which help users get started using the Redshift platform.

sample data, and analyze it using a SQL client.

Get started with the tutorial

Get started with the workshop

Amazon S3

**Getting started with Amazon S3**

Learn how to create your first DataBrew project. You load a sample dataset, run transformations on that dataset, build a recipe to capture those transformations, and run a job to write the transformed data to Amazon S3.

Get started with the guide

**Central storage - Amazon S3 as the data lake storage platform**

Discover how Amazon S3 is an optimal foundation for a data lake because of its virtually unlimited scalability and high durability.

Read the whitepaper

Amazon SageMaker

**How Amazon SageMaker works**

Explore the overview of machine learning and how Amazon SageMaker works.

Explore the guide

**Getting started with Amazon SageMaker**

We show you how to get started using Amazon EMR on EKS by deploying a Spark application on a virtual cluster.

Explore the guide

**Generate machine learning predictions without writing code**

Learn how to use Amazon SageMaker Canvas to build ML models and generate accurate predictions without writing a single line of code.

Get started with the tutorial

# Explore

| Architecture diagrams | Whitepapers | AWS Solutions |
| --- | --- | --- |
| Explore architecture diagrams to help you develop, scale, and test your analytics solutions on AWS. | Explore whitepapers to help you get started, learn best practices, and understand your analytics options. | Explore vetted solutions and architectural guidance for common use cases for analytics services. |
| Explore architecture diagrams | Explore whitepapers | Explore solutions |

# Document history

The following table describes the important changes to this decision guide. For notifications about updates to this guide, you can subscribe to an RSS feed.

| Change | Description | Date |
| --- | --- | --- |
| Initial publication | Guide first published. | November 17, 2023 |