AWS Analytics Services

Created	@February 11, 2024 2:29 PM
Tags	

- Fastest way to get answers from all your data to all your users

What is Data Warehouse

Data warehouse is the process of collecting and storing large amounts of data from various sources into a centralised repository for **the purpose of analysis and reporting** .A data warehouse is a large ,centralized database that is designed to support business intelligence activity such as data analysis ,reporting and decesion making.

benefits include:

- improving data quality
- faster data analysis
- enhanced decision making
- reduced it costs
- data warehouse is not good for transaction its good for reporting and analytics

The are several AWS Analytics services and these include:

- Amazon Athena
- Amazon Redshift
- Amazon EMR
- Amazon CloudSearch
- Amazon Opensearch Service
- Amazon Kinesis
- Amazon QuickSight
- Amazon Data Pipeline
- AWS Glue

- AWS Lake Formation
- Amazon MSK

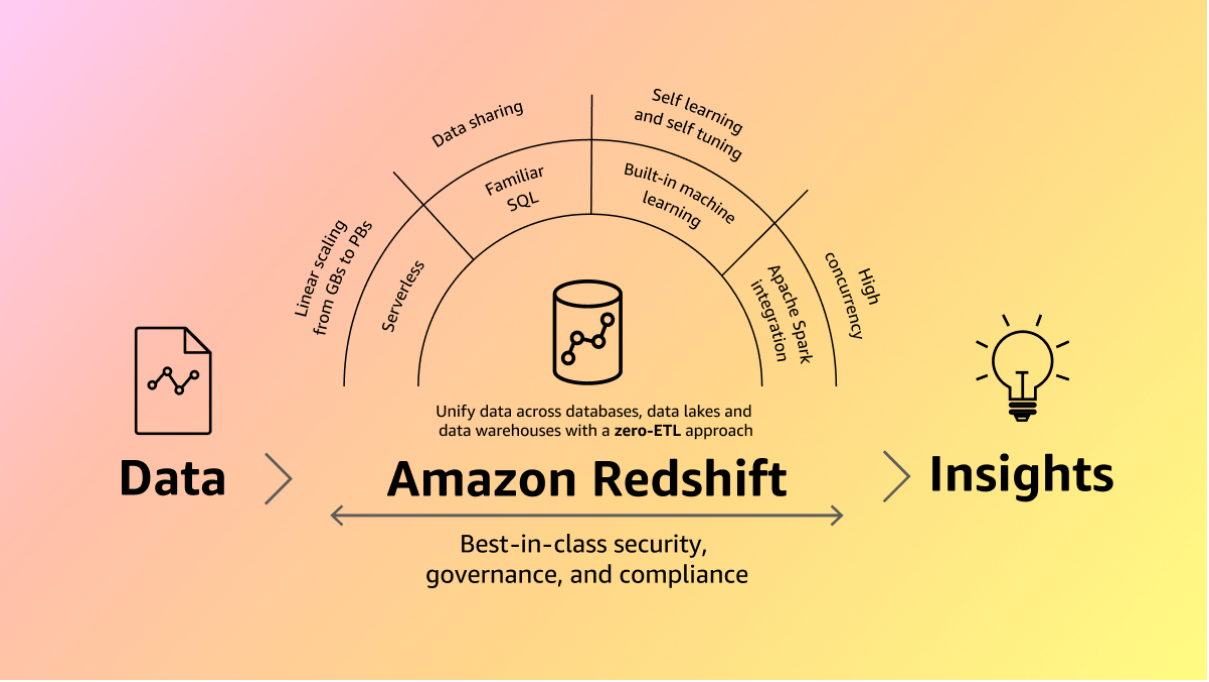
Amazon Redshift

Amazon Redshift ia a fully managed ,petabyte-scale data warehousing service provided by AWS .Redshift is designed to handle large-scale data warehouse and analytics workloads,making it an ideal choice for organizations that need to analyze vast amount of data quickly and efficiently.

- Massively parallel processing (MPP) architecture
- Automatic Compression
- Handles Exabyte-scale data
- Easy integration with S3, Data pipeline, and AWS Glue

How it works

Amazon Redshift uses SQL to analyze structured and semi-structured data across data warehouses, operational databases, and data lakes, using AWS-designed hardware and machine learning to deliver the best price performance at any scale.



Use cases

1. Improve financial and demand forecasts
2. Optimize your business intelligence
3. Accelerate machine learning in SQL
4. Monetize your data
5. Combine your data with third party data sets easily

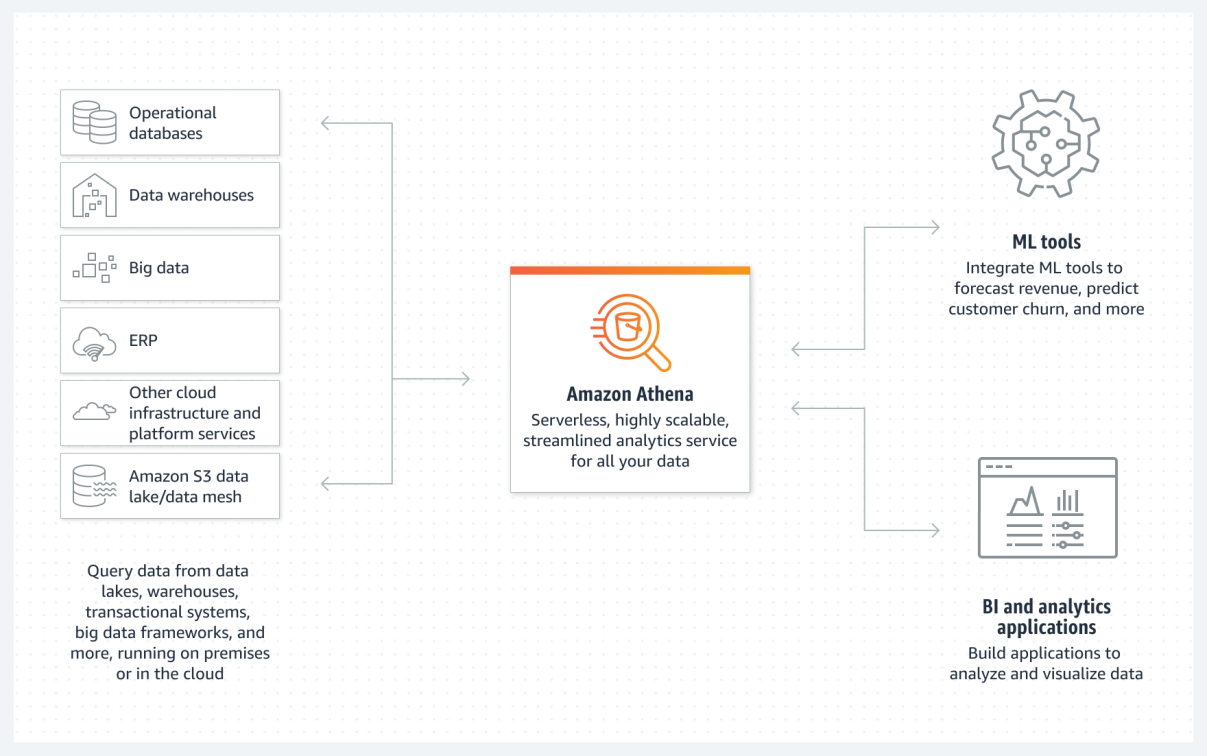
Amazon Athena

Athena is based on the open-source project Apache Presto, which is a distributed SQL query engine with athena ,users can write SQL queries against data storage in S3 and retrive results , regardless of the size of the dataset, athena supports various data formates such as CSV, JSON, Parquet and ORC.

- Athena is a powerful tool for performing ad-hoc queries and analysis on large-scale dataset in S3, without the need for complex infrastructure management.
- Pay per query
- Considered serverless

How it works

Amazon Athena is a serverless, interactive analytics service built on open-source frameworks, supporting open-table and file formats. Athena provides a simplified, flexible way to analyze petabytes of data where it lives. Analyze data or build applications from an Amazon Simple Storage Service (S3) data lake and 30 data sources, including on-premises data sources or other cloud systems using SQL or Python. Athena is built on open-source Trino and Presto engines and Apache Spark frameworks, with no provisioning or configuration effort required.



Use Cases

- Run queries on S3, on premises, or on other clouds
- Prepare data for ML models
- Build distributed big data reconciliation engines
- Perform multicloud analytics

AWS Glue

AWS Glue is built on top of Apache Spark, Which is an open-source distributed computing system. With aws glue, users can define ETL jobs using a drag and drop integration or by writing code in python or scala .AWS Glue crawlers can automatically discover and catalog metadata about storage in various data stores such as Amazon S3, JDBC-compatible databases and amazon Redshift.

- Powerful and flexible ETL service
- Helps to better understand your data
- Crawler
- Glue data catalog

Why AWS Glue?

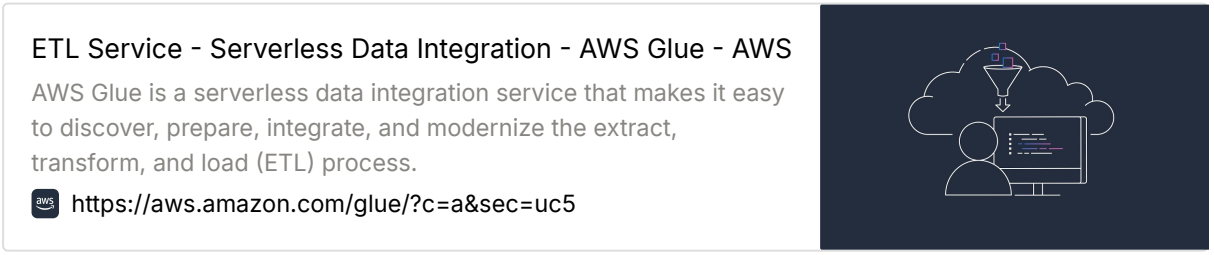
Preparing your data to obtain quality results is the first step in an analytics or ML project. AWS Glue is a serverless data integration service that makes data preparation simpler, faster, and cheaper. You can discover and connect to over 70 diverse data sources, manage your data in a centralized data catalog, and visually create, run, and monitor ETL pipelines to load data into your data lakes.

Benefits of AWS Glue

- Support all workloads
- Scale on demand
- Tailored tools
- All in one

How it works

AWS Glue is a serverless data integration service that makes it easier to discover, prepare, move, and integrate data from multiple sources for analytics, machine learning (ML), and application development.



Use Cases

1. Simplify ETL pipeline development
2. Interactively explore, experiment on, and process data
3. Discover data efficiently
4. Support various processing frameworks and workloads

Amazon Kinesis

It allows users to collect ,process and analyze real-time , streaming data from various sources such as website clickstreams, IOT Devices and social media feeds.

- Analyze real time streaming data
- Support video, audio, application logs, website clickstreams aand IOT

- It can be used for applications such as security monitoring , industrial automation, and live streaming
- Handle terabytes of data per hours
- Data is processed in “shards”.

Use cases

- Stream into data lakes and warehouses
- Boost security
- Build ML streaming applications

There are 4 types of Kinesis service,

1.**Kinesis Video Streams**

Kinesis Video Streams makes it easy to securely stream video from connected devices to AWS for analytics, machine learning (ML), and other processing.

2. **Kinesis Data Streams**

Kinesis Data Streams enables you to build custom applications that process or analyze streaming data for specialized needs.

3. **Kinesis Data Firehose**

Kinesis Data Firehose is the easiest way to load streaming data into data stores and analytics tools.

4. **Kinesis Data Analytics**

Amazon Kinesis Data Analytics is the easiest way to process and analyze real-time, streaming data.

Elastic MapReducer (EMR)

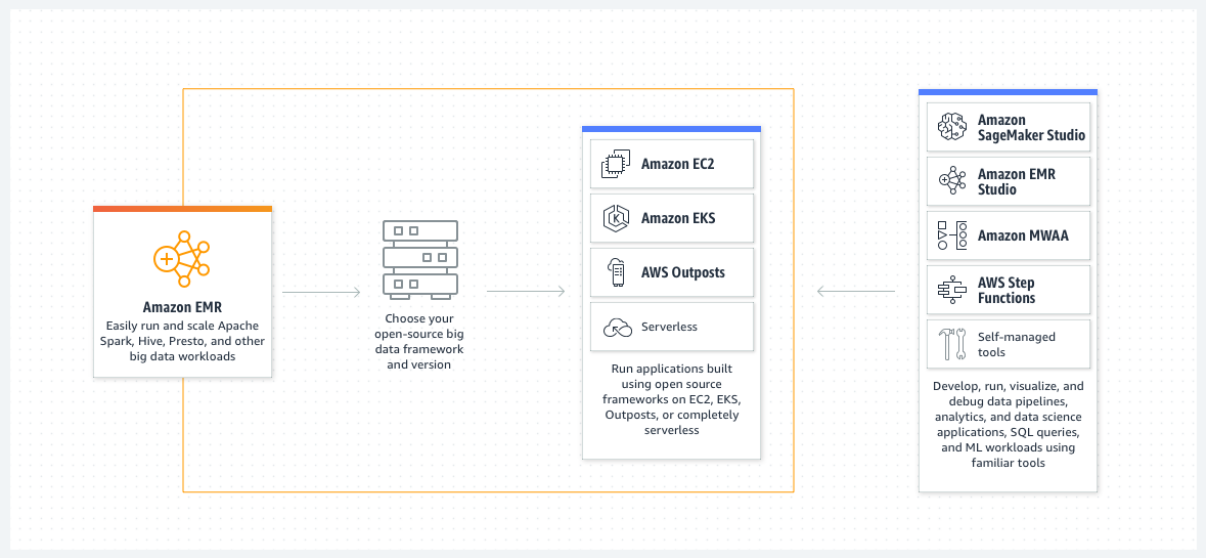
It is designed to make it easy to process large amounts of data using popular big data processing frameworks like Apache Hadoop, Apache Spark and Apache Hive.

- Amazon EMR includes pre-build installations of Hadoop ,spark and other big data processing tools making it easy for users to get started
- Working with big data framework
- Analyse data using Hadoop

- Build scalable data pipelines
- Process real-time data streams

How it works

Amazon EMR is the industry-leading cloud big data solution for petabyte-scale data processing, interactive analytics, and machine learning using open-source frameworks such as Apache Spark, Apache Hive, and Presto.



Use cases:

1. Perform big data analytics
2. Build scalable data pipelines
3. Process real-time data streams
4. Accelerate data science and ML adoption

Data Pipeline

Is a Web service provide by AWS that helps users to easily move data between different AWS services and on-premises data sources, It allows users to define data processing workflows that can automate the movement and transformation of data from various sources to their desired destinations.

- Move data at specific intervals by scheduling
- Moves data based specified conditions

- Sends notifications on success and failure

QuickSight

It provides users with an easy-to-use interface for creating and sharing visualizations, without the need for specialized technical knowledge.

- Interactive visualizations: bar charts, line charts, scatter plots amd more
- Dashboard craeton: layouts , theme and images
- Collaborations and Sharing : Sharing options, such as embedding dashbords in websites or blogs
- Data Exploration: drill down to specific data