AWS Glue

Data Integration

- process of preparing and combining data
- for analytics, machine learning, and application development.

Data Integration

- **It involves multiple tasks, such as**
  - discovering and extracting data from various sources;
  - enriching,
  - cleaning,
  - normalizing, and
  - combining data; and
  - loading and organizing data in databases,
  - data warehouses, and data lakes.

What is AWS Glue?

- **serverless data integration service that makes it easy**
  - to discover,
  - prepare, and
  - combine data
- **for analytics, machine learning, and application development.**

Why AWS Glue?

- **AWS Glue provides both visual and code-based interfaces to make data integration easier.**
- **Users can easily find and access data using the AWS Glue Data Catalog.**
- **Data engineers and ETL (extract, transform, and load) developers can visually create, run, and monitor ETL workflows with a few clicks in AWS Glue Studio.**

Why AWS Glue?

- **Data analysts and data scientists can use AWS Glue DataBrew to visually enrich, clean, and normalize data without writing code.**
- **With AWS Glue Elastic Views, application developers can use familiar Structured Query Language (SQL) to combine and replicate data across different data stores.**

Benefits of AWS Glue

- **Faster data integration**
- **Automate your data integration at scale**
- **No servers to manage**

Use Cases

- **Build event-driven ETL (extract, transform, and load) pipelines**
- **Create a unified catalog to find data across multiple data stores**
- **Create, run, and monitor ETL jobs without coding**

Use Cases

- **Explore data with self-service visual data preparation**
- **Build materialized views to combine and replicate data (in preview)**

Data Pipeline vs Glue

- **AWS Glue is serverless and so there is no infrastructure for developers to manage. Scaling, provisioning, and configuration are fully managed in Glue's Apache Spark environment.**
- **AWS Data Pipeline is not serverless like Glue. It launches and manages the lifecycle of EMR clusters and EC2 instances to execute your jobs.**
- **You can define the pipelines and have more control over the compute resources underlining them.**

Data Pipeline vs Glue

- **AWS Glue provides support for Amazon S3, Amazon RDS, Redshift, SQL, and DynamoDB and also provides built-in transformations.**
- **AWS Data Pipeline allows you to create data transformations through APIs and also through JSON, while only providing support for DynamoDB, SQL, and Redshift.**
- **AWS Glue provides support for Apache Spark framework (Scala and Python) while AWS Data Pipeline supports all the platforms supported by EMR in addition to Shell.**

Data Pipeline vs Glue

- **AWS Glue runs your ETL jobs on its virtual resources in a serverless Apache Spark environment.**
- **AWS Data Pipeline does not restrict to Apache Spark and allows you to make use of other engines like Pig, Hive, etc., thus making it a good choice if your ETL jobs do not require the use of Apache Spark or require the use of multiple engines.**

Demo

Thank you!!!

References

- **https://docs.aws.amazon.com/index.html**