
aws

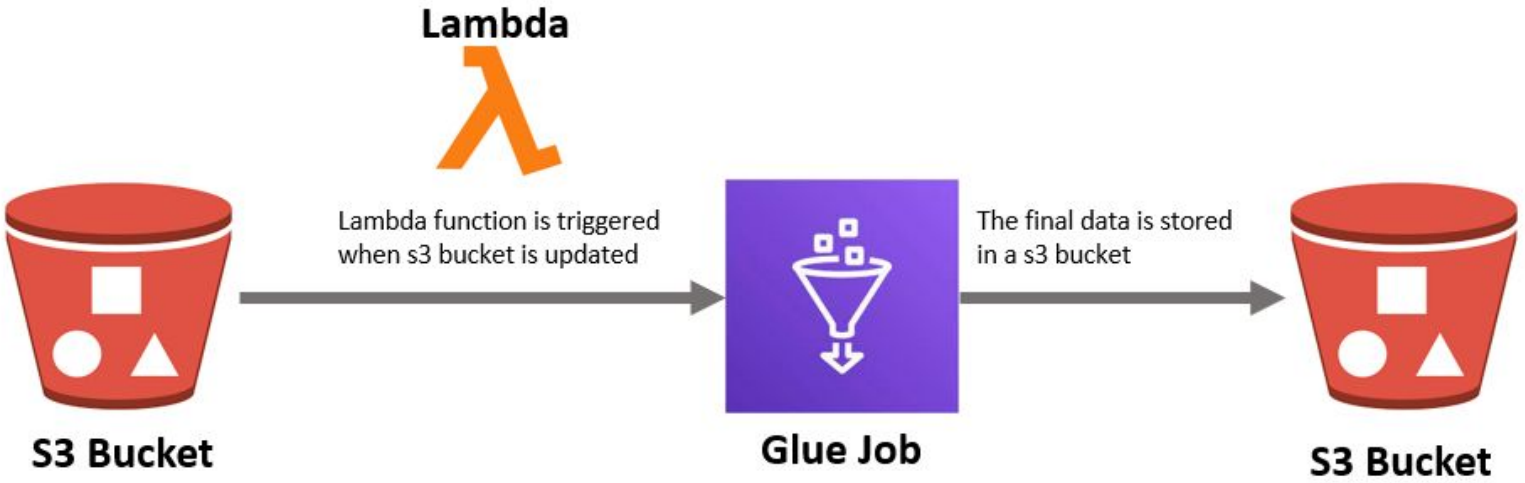
AWS GLUE

Grow Data Skills

AWS Glue is a fully managed extract, transform, and load (ETL) service that makes it simple for users to prepare and load their data for analytics. AWS Glue automates much of the time-consuming data preparation tasks, making it easier for analysts and developers to analyze data, either on AWS or on-premises data stores.

Here are the main components and features of AWS Glue:

- **Data Catalog:** This is your central metadata repository in AWS Glue. The AWS Glue Data Catalog contains table definitions, job definitions, and other control metadata. It replaces the need for a traditional Hive Metastore or any other metadata repository.
- **Crawlers**: These are used to connect to your source data, extract metadata, and create table definitions in the Data Catalog. Crawlers can automatically scan your data in Amazon S3, RDS, Redshift, and more, and then add or update the metadata in the Data Catalog.
- **ETL Jobs:** AWS Glue generates Python or Scala code (based on Apache Spark) for your ETL jobs. You can use the AWS Management Console to design and manage your ETL workflow. Glue will execute these jobs on a fully managed Apache Spark environment.
- **Schedulers**: You can set up recurring ETL jobs with a specified interval, allowing your ETL jobs to handle new data as it arrives.
- **Data Lake and Data Warehouse Integrations:** AWS Glue integrates seamlessly with other AWS services like Amazon S3, Amazon RDS, Amazon Redshift, and more. This makes it a key player in AWS-based data lake architectures and data warehouse integrations.
- **Serverless**: AWS Glue is serverless, meaning you don't need to manage any infrastructure. It automatically provisions and manages the compute resources required to run your ETL jobs.


Lambda
Lambda function is triggered when s3 bucket is updated
The final data is stored in a s3 bucket
S3 Bucket
Glue Job
S3 Bucket