What Is AWS Glue?

06:06 AM

https://docs.aws.amazon.com/glue/latest/dg/what-is-glue.html

AWS Glue is a fully managed **ETL (extract, transform, and load)** service that makes it simple and cost-effective to categorize your data, clean it, enrich it, and move it reliably between various data stores and data streams.

AWS Glue consists of a central metadata repository known as the AWS Glue Data Catalog, an ETL engine that automatically generates Python or Scala code, and a flexible scheduler that handles dependency resolution, job monitoring, and retries.

AWS Glue is serverless, so there's no infrastructure to set up or manage.

AWS Glue is designed to work with semi-structured data. It introduces a component called a dynamic frame, which you can use in your ETL scripts.

A dynamic frame is similar to an Apache Spark dataframe, which is a data abstraction used to organize data into rows and columns, except that each record is self-describing so no schema is required initially.

With dynamic frames, you get schema flexibility and a set of advanced transformations specifically designed for dynamic frames. Y

ou can convert between dynamic frames and Spark dataframes, so that you can take advantage of both AWS Glue and Spark transformations to do the kinds of analysis that you want.

You can use the AWS Glue console to discover data, transform it, and make it available for search and querying.

The console calls the underlying services to orchestrate the work required to transform your data. You can also use the AWS Glue API operations to interface with AWS Glue services.

Edit, debug, and test your Python or Scala Apache Spark ETL code using a familiar development environment.

When Should I Use AWS Glue?

You can use AWS Glue to organize, cleanse, validate, and format data for storage in a data warehouse or data lake.

You can use AWS Glue when you run serverless queries against your Amazon S3 data lake.

AWS Glue can catalog your Amazon Simple Storage Service (Amazon S3) data, making it available for querying with Amazon Athena and Amazon Redshift Spectrum.

AWS Glue Concepts

https://docs.aws.amazon.com/glue/latest/dg/components-key-concepts.html https://towardsdatascience.com/aws-glue-101-all-you-need-to-know-with-a-real-world-example-f34af17b782f

https://docs.aws.amazon.com/glue/latest/ug/tutorial-create-job.html