Modernize Your ETL Data Processing

with AWS Glue To Unleash Better Business Intelligence

AGILISIUM

BIG ON CLOUD. BIG ON DATA.

aws PARTNER Advanced Tier Services

AGILISIUM
BIG ON CLOUD. BIG ON DATA.

In the world of data warehousing, if you need to bring data from multiple different data sources into one, centralized database, you must first:

EXTRACT data from its original source
TRANSFORM data by deduplicating it, combining it, and ensuring quality, to then
LOAD data into the target database

ETL tools enable data integration strategies by allowing companies to gather data from multiple data sources and consolidate it into a single, centralized location. ETL tools also make it possible for different types of data to work together.

Need for Modernization of ETL Platforms

The way businesses obtain, handle, and use data has changed dramatically recently because data sources and consumer interactions have multiplied exponentially. As a result, real-time stream processing and cloud technologies have emerged as the backbone of intelligent decision-making. Thus, the necessity for ETL modernization and how to evaluate ETL systems for a smooth modernization process.

AGILISIUM
BIG ON CLOUD. BIG ON DATA.

Challenges of Traditional ETL Tools

There are several ETL use-cases in which standard ETL solutions are extensively employed.

These challenges include:

- Scalability and flexibility for cutting-edge use cases like machine learning and artificial intelligence
- Time-consuming. Metadata, custom sources, EDW, and data marts must all interact to generate tasks.
- Expensive. Costs of ownership, operation, and maintenance.
- Integration constraints. Connecting to existing infrastructure components is difficult, forcing the development of new use cases.

Problems may also cascade across processes when they are coupled, resulting in data loss. However, a new breed of ETL systems that can run workloads both on-premises and in the cloud are helping enterprises overcome these difficulties.

AGILISIUM
BIG ON CLOUD. BIG ON DATA.

What is AWS Glue?

AWS Glue is a serverless data integration tool for analytics, machine learning, and application development. AWS Glue delivers full data integration features so you can start analyzing and using your data in minutes instead of months.

Prepare and combine data for analytics, machine learning, and application development. It includes wide-ranging operations like data discovery and extraction, enrichment, cleansing, normalization, combination, and data loading and storage in databases, warehouses, and data lakes. These actions are often performed by individuals who utilize various products.

How Does It Work?

For your organization, AWS Glue has several features and benefits. The AWS Management Console makes it shockingly simple to set up and run an ETL operation.

This is how it works:

Add information to S3 files by defining crawlers.

You may do a scan regularly or just when a particular event happens. Using AWS Glue and Python Glue, you can build an ETL pipeline that runs on an AWS Spark cluster infrastructure, and you'll only be charged when the operation is completed.

You can keep your data distinct from your processing engines with AWS Glue. A wide range of processing engines may access the information. In some instances, an API Gateway may expose and route all catalog requests via it.

Features of AWS Glue

There is no need to re-run a method on the same data to debug and measure performance in the same S3 bucket.

Here are some cool AWS features and concepts

Control and Logging

AWS-Glue logs are forwarded to Amazon CloudWatch by default

.

Prompts

AWS Glue offers enhanced monitoring options in addition to standard logs. Verify the selections during or after each task.

Unfortunately, python shell jobs can't access monitoring. Automation

AWS Glue Supports ETL Automation Through Triggers and Workflows

Using AWS Glue triggers, you can build a fundamental process and crawler chain.

WebGL Methods and Workflows

With numerous crawlers, tasks, and triggers, workflows may help build and visualize complex ETL processes.

Each process oversees its components. In addition, AWS Glue connects with AWS services like Lambda and AWS Step Functions.

Tagging Jobs

When doing ETL, AWS Glue leverages job bookmarks to prevent reprocessing.

AWS Glue keeps track of ETL data by saving task run status information. The job bookmark stores the state.

Save state using job bookmarks and prevent reprocessing outdated data. See also Job bookmarks for data tracking.

AGILISIUM
BIG ON CLOUD. BIG ON DATA.

Advantages of Migrating to AWS Glue

AWS Glue natively supports data stored in Amazon Aurora, Amazon RDS engines, Amazon Redshift, Amazon S3, as well as conventional database engines and Amazon VPC. This reduces onboarding hassle.

AWS Glue is serverless, so there are no computing resources to maintain. It also manages resource provisioning, setup, and scalability for your ETL processes in a fully managed, scale-out Apache Spark environment. This saves money since you just pay for the resources utilized throughout your tasks.

More Power: AWS Glue simplifies the creation, maintenance, and execution of ETL tasks. So it snoops through your databases, looking for patterns in the data. It also creates code for data conversions and loading procedures.

When To Use AWS Glue

To help you decide where to utilize AWS Glue, below are some examples of use cases and how AWS Glue may help you:

Amazon S3 Data Lake Queries:

Want to create your own Amazon S3 data lake? AWS Glue can make it happen instantly by making all your data analytics-ready without relocating it.

Data Warehouse Log Analysis:

Your data warehouse's semi-structured data may be processed quickly using AWS Glue. It develops data structure, ETL code to convert, flatten, and enhance data and regularly fills your data warehouse.

Unified Data View Across Multiple Data Stores:

It enables you to search across various AWS data sets without transferring them. It unifies your data and makes it searchable and queryable utilizing Amazon Athena, Amazon EMR, and Amazon Redshift Spectrum.

ETL Pipelines with Events:

AWS Glue can conduct ETL processes depending on events, such as new data sets. For example, you may utilize an AWS Lambda function to perform ETL operations whenever new data in Amazon S3 becomes available. You can also add this new dataset to your ETL processes in the AWS Glue Data Catalog.

Businesses have depended on ETL systems for decades to acquire a clearer perspective on their data. They're still a vital part of data integration tool kits. So there you have it, an overview of AWS Glue's data cataloguing and ETL automation capabilities. In addition, Agilisium Consulting is a certified AWS Advanced Consulting Partner who offers No-Cost Assessment program with risk-free mitigation plan. Contact us today to support you ETL Modernization needs, and our AWS experts will readily help.

sales@agilisium.com