
AWS Glue Samples

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. This repository has samples that demonstrate various aspects of the AWS Glue service, as well as various AWS Glue utilities.

You can find the AWS Glue open-source Python libraries in a separate repository at: [awslabs/aws-glue-lib](https://github.com/aws-labs/aws-glue-lib).

Getting Started

- **Getting Started with AWS Glue**
Helps you get started using AWS Glue.
- **FAQ and How-to**
Helps you get started using the many ETL capabilities of AWS Glue, and answers some of the more common questions people have.

Workshops

- **AWS Glue Learning Series**
In this comprehensive series, you'll learn everything from the basics of Glue to advanced optimization techniques.

Tutorials

General

- **Writing an AWS Glue for Spark script**
Introduces the process of writing AWS Glue scripts.
- **Detect and process sensitive data using AWS Glue Studio**
Guides you to create an AWS Glue job that identifies sensitive data at the row level, and create a custom identification pattern to identify case-specific entities.

-
- Enable self-service visual data integration and analysis for fund performance using AWS Glue Studio and Amazon QuickSight

Demonstrates how AWS Glue Studio helps you perform near-real-time analytics, and how to build visualizations and quickly get business insights using Amazon QuickSight.

- Stream data from relational databases to Amazon Redshift with upserts using AWS Glue streaming jobs

Guides you to setup Change data capture (CDC) from relational databases to Amazon Redshift with enriching data using Glue Streaming job.

- AWS Glue streaming application to process Amazon MSK data using AWS Glue Schema Registry
- Shows how to use a combination of Amazon MSK, the AWS Glue Schema Registry, AWS Glue streaming ETL jobs, and Amazon S3 to create a robust and reliable real-time data processing platform.

Data migration

- Implement vertical partitioning in Amazon DynamoDB using AWS Glue

Guides you to use AWS Glue to perform vertical partitioning of JSON documents when migrating document data from Amazon S3 to Amazon DynamoDB.

- Migrate terabytes of data quickly from Google Cloud to Amazon S3 with AWS Glue Connector for Google BigQuery

Guides you to use AWS Glue to build an optimized ETL process to migrate a large and complex dataset from Google BigQuery storage into Amazon S3 in Parquet format.

- Migrate from Google BigQuery to Amazon Redshift using AWS Glue and Custom Auto Loader Framework

Guides you to use AWS Glue to migrate from Google BigQuery to Amazon Redshift.

- Migrate from Snowflake to Amazon Redshift using AWS Glue Python shell

Guides you to use AWS Glue Python shell jobs to migrate from Snowflake to Amazon Redshift.

- Compose your ETL jobs for MongoDB Atlas with AWS Glue

Guides you to use AWS Glue to process data into MongoDB Atlas.

Open Table Format

- Introducing native support for Apache Hudi, Delta Lake, and Apache Iceberg on AWS Glue for Apache Spark
 - Part 1: Getting Started
 - Part 2: AWS Glue Studio Visual Editor

This series of posts demonstrate how you can use Apache Hudi, Delta Lake, and Apache Iceberg on Glue Studio notebook and Glue Studio Visual Editor.

- Implement a CDC-based UPSERT in a data lake using Apache Iceberg and AWS Glue
Guides you to setup Change data capture (CDC) from relational databases to Iceberg-based data lakes using Glue job.
- Implement slowly changing dimensions in a data lake using AWS Glue and Delta
Demonstrates how to identify the changed data for a semi-structured source (JSON) and capture the full historical data changes (SCD Type 2) and store them in an S3 data lake.

Development, Test, and CI/CD

- Develop and test AWS Glue version 3.0 and 4.0 jobs locally using a Docker container
Gives you an instruction to develop and test Glue scripts locally using a Docker container. This tutorial includes different methods like `spark-submit`, REPL shell, unit test using `pytest`, notebook experience on JupyterLab, and local IDE experience using Visual Studio Code.
- Build, Test and Deploy ETL solutions using AWS Glue and AWS CDK based CI/CD pipelines
Gives you an instruction to build CI/CD pipelines for AWS Glue components using AWS CDK.

Cost and Performance

- Monitor and optimize cost on AWS Glue for Apache Spark
Demonstrates best practices to monitor and optimize cost on AWS Glue for Apache Spark. This tutorial also includes a template to set up automated mechanism to collect and publish DPU Hours metrics in CloudWatch.
- Best practices to optimize cost and performance for AWS Glue streaming ETL jobs
Demonstrates best practices to optimize cost and performance for AWS Glue streaming ETL jobs.

Glue for Ray

- **Introducing AWS Glue for Ray: Scaling your data integration workloads using Python**
Provides an introduction to AWS Glue for Ray and shows you how to start using Ray to distribute your Python workloads.
- **Scale AWS SDK for pandas workloads with AWS Glue for Ray**
Shows you how to use pandas to connect to AWS data and analytics services and manipulate data at scale by running on an AWS Glue for Ray job.
- **Advanced patterns with AWS SDK for pandas on AWS Glue for Ray**
Shows how to use some of these APIs in an AWS Glue for Ray job, namely querying with S3 Select, writing to and reading from a DynamoDB table, and writing to a Timestream table.

Glue Data Catalog

- **Get started managing partitions for Amazon S3 tables backed by the AWS Glue Data Catalog**
Covers basic methodologies for managing partitions for Amazon S3 tables in Glue Data Catalog.
- **Improve query performance using AWS Glue partition indexes**
Demonstrates how to utilize partition indexes, and discusses the benefit you can get with partition indexes when working with highly partitioned data.

Glue Crawler

- **Adding an AWS Glue crawler**
Provides an introduction to AWS Glue crawler.
- **Efficiently crawl your data lake and improve data access with an AWS Glue crawler using partition indexes**
Describes how to create partition indexes with an AWS Glue crawler and compare the query performance improvement when accessing the crawled data with and without a partition index from Athena.
- **AWS Glue crawlers support cross-account crawling to support data mesh architecture**
Walks through the creation of a simplified data mesh architecture that shows how to use an AWS Glue crawler with Lake Formation to automate bringing changes from data producer domains to data consumers while maintaining centralized governance.

Glue Data Quality

- Getting started with AWS Glue Data Quality from the AWS Glue Data Catalog
Provides an introduction to AWS Glue Data Quality.
- Getting started with AWS Glue Data Quality for ETL Pipelines
Shows how to create an AWS Glue job that measures and monitors the data quality of a data pipeline.
- Set up advanced rules to validate quality of multiple datasets with AWS Glue Data Quality
Demonstrates the advanced data quality checks that you can typically perform when bringing data from a database to an Amazon S3 data lake.
- Set up alerts and orchestrate data quality rules with AWS Glue Data Quality
Explains how to set up alerts and orchestrate data quality rules with AWS Glue Data Quality.
- Visualize data quality scores and metrics generated by AWS Glue Data Quality
Explains how to build dashboards to measure and monitor your data quality.

Glue ETL Code Examples

You can run these sample job scripts on any of AWS Glue ETL jobs, container, or local environment.

- Join and Relationalize Data in S3
This sample ETL script shows you how to use AWS Glue to load, transform, and rewrite data in AWS S3 so that it can easily and efficiently be queried and analyzed.
- Clean and Process
This sample ETL script shows you how to take advantage of both Spark and AWS Glue features to clean and transform data for efficient analysis.
- The `resolveChoice` Method
This sample explores all four of the ways you can resolve choice types in a dataset using DynamicFrame's `resolveChoice` method.
- Converting character encoding
This sample ETL script shows you how to use AWS Glue job to convert character encoding.
- Notebook using open data lake formats

The sample iPython notebook files show you how to use open data lake formats; Apache Hudi, Delta Lake, and Apache Iceberg on AWS Glue Interactive Sessions and AWS Glue Studio Notebook.

- Blueprint examples

The sample Glue Blueprints show you how to implement blueprints addressing common use-cases in ETL. The samples are located under `aws-glue-blueprint-libs` repository.

Utilities

- Hive metastore migration

This utility can help you migrate your Hive metastore to the AWS Glue Data Catalog.

- Crawler undo and redo

These scripts can undo or redo the results of a crawl under some circumstances.

- Spark UI

You can use this Dockerfile to run Spark history server in your container. See details: [Launching the Spark History Server and Viewing the Spark UI Using Docker](#)

- use only IAM access controls

AWS Lake Formation applies its own permission model when you access data in Amazon S3 and metadata in AWS Glue Data Catalog through use of Amazon EMR, Amazon Athena and so on. If you currently use Lake Formation and instead would like to use only IAM Access controls, this tool enables you to achieve it.

- Glue Resource Sync Utility

This utility enables you to synchronize your AWS Glue resources (jobs, databases, tables, and partitions) from one environment (region, account) to another.

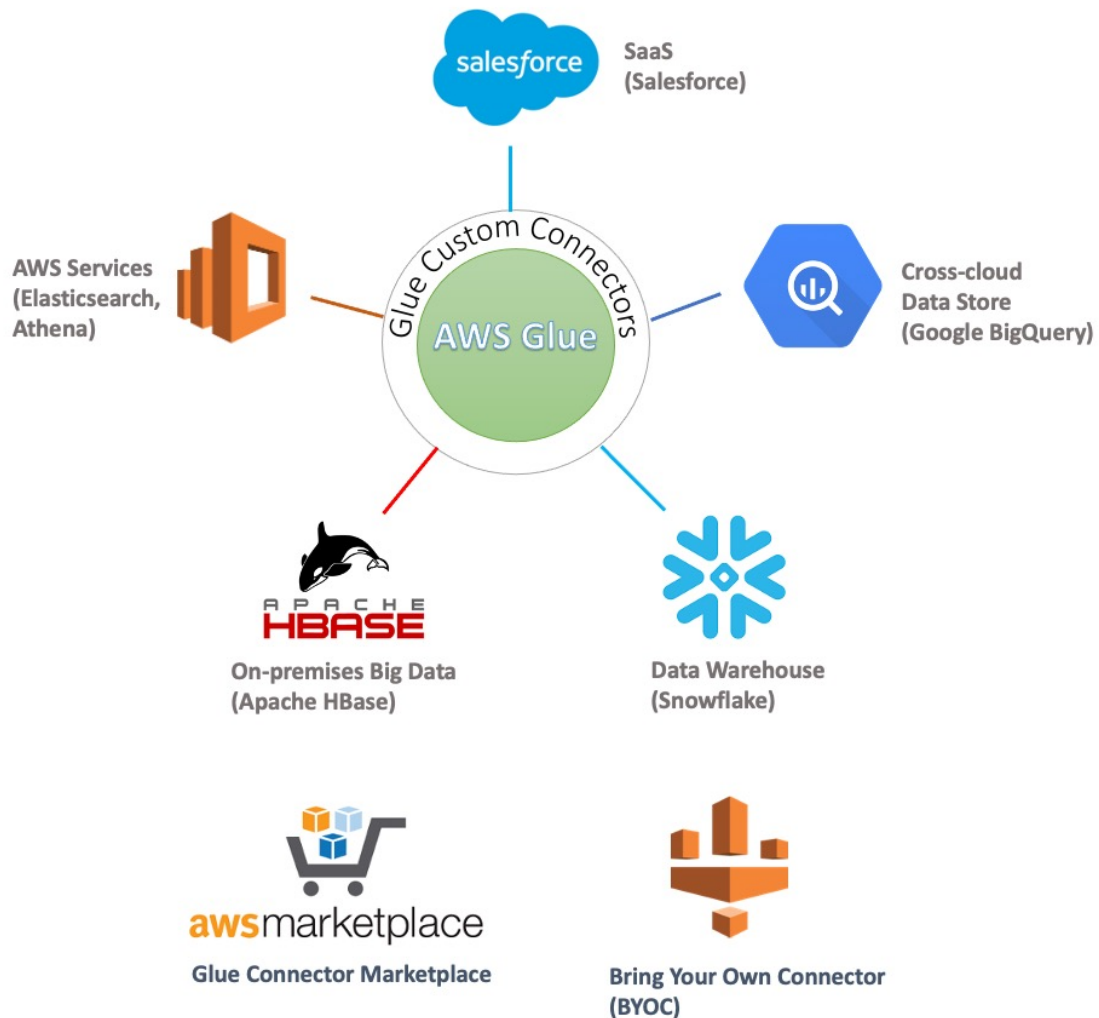
- Glue Job Version Deprecation Checker

This command line utility helps you to identify the target Glue jobs which will be deprecated per AWS Glue version support policy.

Glue Custom Connectors

AWS Glue provides built-in support for the most commonly used data stores such as Amazon Redshift, MySQL, MongoDB. Powered by Glue ETL Custom Connector, you can subscribe a third-party connector

from AWS Marketplace or build your own connector to connect to data stores that are not natively supported.



- **Development**

Development guide with examples of connectors with simple, intermediate, and advanced functionalities. These examples demonstrate how to implement Glue Custom Connectors based on Spark Data Source or Amazon Athena Federated Query interfaces and plug them into Glue Spark runtime.

- **Local Validation Tests**

This user guide describes validation tests that you can run locally on your laptop to integrate your connector with Glue Spark runtime.

- Validation

This user guide shows how to validate connectors with Glue Spark runtime in a Glue job system before deploying them for your workloads.

- Glue Spark Script Examples

Python scripts examples to use Spark, Amazon Athena and JDBC connectors with Glue Spark runtime.

- Create and Publish Glue Connector to AWS Marketplace

If you would like to partner or publish your Glue custom connector to AWS Marketplace, please refer to this guide and reach out to us at glue-connectors@amazon.com for further details on your connector.

License Summary

This sample code is made available under the MIT-0 license. See the LICENSE file.