DATA INTEGRATION & FEDERATION

Analytics in 15

Opensource Table Formats with AWS Glue and Amazon EMR

Priyanka Chaudhary

Big Data Architect
Amazon Web Services

Agenda

- AWS Glue Overview
- Amazon EMR Overview
- Opensource Table Formats
- Demo
- Additional Resources

Introduction to AWS Services

AWS Glue

- Discover, prepare, and integrate all your data at any scale

All-in-one data integration service

Cost effective, serverless, and scalable

Tailored tools to support all data users

Support all workloads in one place

4

Amazon EMR

- Easily Run Spark, Hive, Presto, HBase, Flink, and more big data apps on AWS

Latest versions

Updated with latest open source frameworks within 60 days

Support for popular OSS like Spark, Flink, Hudi, Iceberg etc.

Great performance at lower cost

Spark workloads run up to 3x faster compared to open source

50–80% reduction in costs with EC2 Spot, EC2 Savings Plan, Reserved Instances

Per-second billing for flexibility

Use S3 storage

Process data in S3 securely with high performance using the EMRFS connector

Scale Compute and Storage independent of each other

Multiple deployment options

Fully managed, choose multiple deployment options based on your need

Opensource Table Formats

AWS Glue Native support for Hudi, Delta Lake and Iceberg

Simplify incremental data processing in data lakes built on Amazon S3

- Built in support to read and write (insert, update and delete)
- No setup, minimal configuration and easy to deploy
- Latest framework versions available with Glue v4.0

Transactional data lakes

- Choice of framework for each workload

EMR 6.10 includes:

Apache Hudi 0.12.2

Apache Iceberg 1.1.0

OSS Delta Lake 2.2.0

Opensource Table Formats Key Features

Apache hudi

- Open table format to build streaming data lakes with incremental data pipelines
- Improve the lookup performance in file index and query latency with data skipping
- Index columns in the background without affecting writes
- Improved Schema evolution support

Apache iceberg

- Open table format for huge analytic datasets
- Time travel support with Spark SQL and Trino SQL
- Merge on Read (MoR) support
- Optimistic concurrency with AWS Glue Data Catalog
- Disaster recovery with S3 access points

Delta Lake

- Open-source storage framework that enables building a Lakehouse Architecture
- Provides ACID transactions and SQL support for creates, inserts, merges, updates, and deletes.
- Uses Optimistic Control to handle concurrency, since the majority of data lake operations append data to a time-ordered partition and won’t conflict.

Demo

Additional Resources

- Blog: https://aws.amazon.com/blogs/big-data/part-1-integrate-apache-hudi-delta-lake-apache-iceberg-datasets-at-scale-aws-glue-studio-notebook/
- Blog: https://aws.amazon.com/blogs/big-data/part-2-integrate-apache-hudi-delta-lake-apache-iceberg-dataset-at-scale-using-aws-glue-studio-visual-editor/
- Workshop: https://catalog.us-east-1.prod.workshops.aws/workshops/65bfc724-91ec-4c89-949a-f78304adf3ff/en-US/150-workshop-dataset

Thank you!

To learn more, please visit:

https://aws.amazon.com/big-data/datalakes-and-analytics/