Introduction to Apache Spark and AWS Glue

Key Concepts, Components, and Governance Practices

Apache Spark Overview

- What is Spark?
- Apache Spark is an open-source, distributed computing engine designed for big data processing and analytics.
- Provides fast, in-memory computations for large-scale data workloads.

Lazy Evaluation Architecture

Transformations (like map, filter) are not executed immediately; instead, Spark builds a logical execution plan.

Actions (like collect, count) trigger the actual computation.

This approach optimizes query execution and minimizes unnecessary data processing.

Main Components of Spark

Spark Core: Basic functions like task scheduling and memory management.

Spark SQL: Module for structured data processing using DataFrames and SQL queries.

Spark Streaming: Real-time data processing engine.

MLlib: Machine learning library for scalable algorithms.

GraphX: Library for graph computation and analytics.

AWS Glue Data Catalog

- Central Metadata Repository
- Acts as a unified store for metadata about data assets across your organization.
- Integrates with various AWS services (S3, Redshift, RDS, Athena, etc.).
- Supports schema, partitioning, and data location details for all integrated datasets.

AWS Glue Job Types

- Spark (Standard) Jobs: Distributed ETL processing using Spark engine.
- Python Shell Jobs: Lightweight, single-threaded scripts for simple tasks or custom logic.
- Streaming ETL: Real-time ETL jobs that process streaming data, scaled by DPUs (Data Processing Units).
- Built-in Transformations & Data Quality Rules:
- Predefined transformations for cleaning, joining, and transforming data.
- Data quality rules help ensure data validity and integrity.

Glue Crawlers

- Configuring Data Sources: Connect to S3, RDS, Redshift, JDBC sources, and more.
- Scheduling Crawlers: Automate metadata extraction by scheduling crawlers to run at regular intervals.
- Schema Evolution & Partitioning:
- Crawlers detect changes in schemas and update the Data Catalog automatically.
- Support for partitioned datasets to optimize query performance.

Glue Data Catalog as a Governance Tool

- Technical Metadata:
- Includes schema details, partitions, data location, and format.

Business Metadata:

Definitions, descriptions, data ownership, and stewardship information.

Operational Metadata:

Data quality scores, job execution lineage, and audit logs.

Data Discovery for Governance

- How Users Find Trustworthy Data:
- Search and browse features in the Data Catalog enable users to locate relevant datasets.
- Metadata (technical, business, operational) helps users evaluate data reliability and suitability for their use cases.

- Governance policies and data quality metrics further support trust in discovered data assets.