ETL pipeline on YouTube data using Athena, Glue and Lambda

Business Overview

Many problems exist when deploying or transferring analytics to the cloud. Differences in features between on-premises and cloud data platforms, security, and governance are all technical concerns. The danger of moving on-premises data into the cloud has prompted organizations to limit cloud analytics initiatives, especially in regulated industries where data protection is crucial. Cloud-based safe Data Lake solutions aid in the development of rich analytics on data while classifying it into several storage phases, such as raw, cleansed, and analytical. This project aims to securely manage, streamline, and perform analysis on the structured and semi-structured YouTube videos data based on the video categories and the trending metrics.

Data Pipeline

A data pipeline is a technique for transferring data from one system to another. The data may or may not be updated, and it may be handled in real-time (or streaming) rather than in batches. The data pipeline encompasses everything from harvesting or acquiring data using various methods to storing raw data, cleaning, validating, and transforming data into a query-worthy format, displaying KPIs, and managing the above process.

Dataset Description

This Kaggle dataset contains statistics (CSV files) on daily popular YouTube videos over the course of many months. There are up to 200 trending videos published every day for many locations. The data for each region is in its own file. The video title, channel title, publication time, tags, views, likes and dislikes, description, and comment count are among the items included in the data. A category_id field, which differs by area, is also included in the JSON file linked to the region.

Tech Stack:

➔ Languages SQL, Python3
➔ Services AWS S3, AWS Glue, QuickSight, AWS Lambda, AWS Athena, AWS IAM

Amazon S3

Amazon S3 is an object storage service that provides manufacturing scalability, data availability, security, and performance. Users may save and retrieve any quantity of data using Amazon S3 at any time and from any location.

AWS IAM
This is nothing but identity and access management which enables us to manage access to AWS services and resources securely. One can create and manage AWS users and groups, use permissions to allow and deny their access to AWS resources. It is a feature of AWS with no additional charge.

QuickSight
Amazon QuickSight is a scalable, serverless, embeddable, machine learning powered business intelligence (BI) service built for the cloud. It is the first BI service to offer pay-per-session pricing, where you only pay when your users access their dashboards or reports, making it cost-effective for large-scale deployments. It can connect to various sources like Redshift, S3, Dynamo, RDS, files like JSON, text, CSV, TSV, Jira, Salesforce, and on-premises oracle SQL-server.

AWS Glue
A serverless data integration service makes it easy to discover, prepare, and combine data for analytics, machine learning, and application development. It runs Spark/Python code without managing Infrastructure at a nominal cost. You pay only during the run time of the job. Also, you pay storage costs for Data Catalog objects. Tables may be added to the AWS Glue Data Catalog using a crawler. The majority of AWS Glue users employ this strategy. In a single run, a crawler can crawl numerous data repositories. The crawler adds or modifies one or more tables in your Data Catalog after it's finished.

AWS Lambda
Lambda is a computing service that allows programmers to run code without having to create or manage servers. Lambda executes the code on high-availability computing infrastructure and manages all aspects of it, including server and operating system maintenance, capacity provisioning and automated scaling, code monitoring, and logging. Lambda allows you to run code for almost any form of application or backend service.

AWS Athena
Athena is an interactive query service for S3 in which there is no need to load data it stays in S3. It is serverless and supports many data formats e.g CSV, JSON, ORC, Parquet, AVRO.

Key Takeaways
- Understanding the project Overview and Architecture
- Understanding ETL on Big Data
- Introduction to Staging and Data Lake

- Creating IAM Roles and Policies
- Creating Lambda Functions
- Setting up Glue Jobs for ETL
- Using Glue Crawler and Glue Studio
- Creating Glue Data Catalog
- Converting JSON to Parquet format
- Performing Data Transformations and Joins
- Visualizing in QuickSight

Architecture