Big Data Project to Learn AWS Athena

Business Overview

Amazon Athena is a query service that allows you to analyze data directly in Amazon S3 using conventional SQL. Using a few clicks in the AWS Management Console, you can aim Athena at Amazon S3 data and start running ad-hoc searches with traditional SQL in seconds. Because Athena is serverless, you don't have to worry about setting up or maintaining infrastructure, and you just pay for the queries you perform. Even with big datasets and sophisticated queries, Athena grows automatically while processing queries in parallel, resulting in fast responses. In addition, Athena supports a variety of data formats, including CSV, JSON, ORC, Parquet, and AVRO.

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

A Covid-19 dataset will be used for this project's demo purpose, which includes timestamps, posts, and comments related to Covid.

➔ Languages-
- SQL, Python3

➔ Services -
- AWS S3, AWS Glue, AWS Athena, Amazon CloudWatch

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 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.

Key Takeaways

- Understanding the project Overview and Architecture
- Table creation using Glue Crawler
- Table creation using CTAS
- Table creation using DDL
- Understanding Athena Partitioning
- Understanding Athena Bucketing
- Exploring Joins in Athena
- Optimizations in Athena
- Creating Athena Workgroup
- Understanding Athena File Formats
- Understanding Athena Pricing

Dataset link- https://www.kaggle.com/datasets/pavellexyr/the-reddit-covid-dataset