Hadoop Project-Analysis of Yelp Dataset using Hadoop Hive

Business Overview

Big Data is the collection of huge datasets of semi-structured and unstructured data, generated by the high-performance heterogeneous group of devices ranging from social networks to scientific computing applications. Companies have the potential to gather large volumes of data, and they must guarantee that the data is in a highly useable shape by the time it reaches data scientists and analysts. Data engineering is the profession of creating and constructing systems for gathering, storing, and analyzing large amounts of data. It is a vast field with applications in almost every sector.

Apache Hadoop is a Big Data technology that enables the distributed processing of massive data volumes across computer clusters using simple programming concepts. It is intended to grow from a single server to thousands of computers, each supplying local computing and storage.

Yelp is a community review site and an American multinational firm based in San Francisco, California. It publishes crowd-sourced reviews of local businesses as well as the online reservation service Yelp Reservations. Yelp has made a portion of their data available in order to launch a new activity called the Yelp Dataset Challenge, which allows anyone to do research or analysis to find what insights are buried in their data. Due to the bulk of the data, this project only selects a subset of Yelp data. User and Review dataset is considered for this session.

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.

Tech Stack

➔

Language: HQL

➔

Services: AWS EMR, Hive, HDFS, AWS S3

AWS EMR

Amazon EMR is a managed cluster platform that makes it easier to use big data frameworks like Apache Hadoop and Apache Spark to handle and analyze large volumes of data on AWS. You may process data for analytics and business intelligence tasks using these frameworks and related open-source projects. Amazon EMR also allows you to convert and transport huge volumes of data across AWS data storage and databases, such as Amazon S3 and Amazon DynamoDB.

Hive

Apache Hive is a fault-tolerant distributed data warehousing solution that enables massive-scale analytics. Using SQL, Hive allows users to read, write, and manage petabytes of data.

Hive is based on Apache Hadoop, an open-source system for storing and processing massive information. As a result, Hive is tightly linked with Hadoop and is built to handle petabytes of data fast. The ability to query massive datasets with a SQL-like interface, using Apache Tez or MapReduce, distinguishes Hive.

Key Takeaways

- Understanding Project overview
- Introduction to Big Data
- Overview of Hadoop ecosystem
- Understanding Hive concepts
- Understanding the dataset
- Implementing Hive table operations
- Creating static and dynamic Partitioning
- Creating Hive Buckets
- Understanding different file formats in Hive
- Using Complex Hive Functions in Hive
- Launching EMR cluster in AWS

NOTE:

- EMR cluster master node can be connected via OpenSSH in Windows OS and files can be transferred using scp command
- For connection through PuTTY, the PEM file has to be converted to PPK using PuTTYgen