DATA ENGINEERING ON YELP DATASET USING HADOOP & HIVE

Hadoop - HDFS & MR

- **Hadoop - Hadoop is an open source,Java based programming framework that manages data processing and storage for big data applications in a distributed computing system.**
- **Features of Hadoop - Flexibility,Scalability,Fault-proof,Cost-effective**
- **Hadoop vs RDBMS - Hadoop doesn't support real time data processing (OLTP),**

it is designed to support large scale batch processing work loads (OLAP) where as RDBMS supports real-time data processing.

components of HDFS

- NAME NODE
- DATA NODE
- Secondary NameNode
- STANDBY Namenode

YARN - Yet Another Resource Negotitator

- YARN also follows the Master-Slave architecture.
- Master - ResourceManager
- Slave - NodeManager

Key aspects of HDFS

- REPLICATION IN HDFS - In HDFS,Default replication factor is 3
- SPLITTING OF DATA - In HDFS,Default block size is 128MB.
- File Operations - Reading from in file , Write to a file

MAP-REDUCE

- MAP - Data parallel model is used in MAP phase.
- Reduce - Inverse-tree parallel model for aggregating results in Reduce phase

HIVE

- Hive is a data warehouse software
- Hive vs RDBMS - Hive - Write once, Read many times ; RDBMS - Read and Write many times.
- Uses of Hive-Metastore - It stores all the information about Hive tables, in a central repository.
- HIVE Data Types - Primitive Data type , Complex Data type
- Primitive Data type - String,Int,float,double,Timestamp,Binary
- Complex Data type – STRUCT(Object),MAP(Key-Value),ARRAY(Indices)
- Internal vs External tables
- External table - External tables is used when data to remain stored on the HDFS even after dropping tables

  because Hive does not delete the data stored outside (of the Hive database).

- Internal table - In internal table, the data is temporary. So, when the Hive table is dropped, the data stored

  in the internal table is deleted along with the metadata.

YELP Dataset Overview

- The Challange Dataset: 2.7M reviews and 649K tips by users for 86K business,566K business attributes,
- (e.g) hours,parking,availability,ambience. Social network of 687K users for a
- total of 4.2M social edges Agregated check-ins over time for each of the 86K businesses
- 200,000 pictures from the included businesses
- Intrested areas of research - Cultural Trends,Location Mining and Urban panning,Seasonal Trends
- Infer Categories,Natural language ProcessingNLP,Changepoints and Events,Social Graph Mining
- Intrested areas of research - Cultural Trends,Location Mining and Urban panning,Seasonal Trends
- Infer Categories,Natural language ProcessingNLP,Changepoints and Events,Social Graph Mining
- Dataset's - Domain,Business,Review,User,check-in,tip,photo
- User and Review dataset considered for this session.

YELP Data modeling - Table creations

- Structure of user.Json
- Structure of review.Json

Basic Queries in Hive and few UDF explanation

- Query to give the total number of records, total number of unique user, min and avg star values
- Finding avg review each reviwer given
- Average stars given by reviewer
- Join Statement - Joining user and review table's

Hive Table Partitioning

- Partitions are used to make queries faster by dividing the tables into smaller parts using partition key columns.

- Types of Partitioning - Static and Dynamic Partitioning
- STATIC PARTIONING - Insert input data files individually into a partition table is Static Partition.
- DYNAMIC PARTIONING - Single insert to partition table is known as dynamic partition.

Usually dynamic partition load the data from non partitioned table

- Property set - set hive.exec.dynamic.partition=true;

set hive.exec.dynamic.partition.mode = nonstrict;

Bucketing in Hive tables

- Bucketing also divides the data into smaller and more manageable parts based on hash value of a column value.
- Property set - set hive.enforce.bucketing=true;

File formats in Hive

  - ORC - Optimized Row Columnar - good compression ratio, avoid unwanted data seek using strips((collection of column data) to store data.

Uses the concept of column major order. Contents are stored as Binary.

- Syntax - stored as orc

location '<<Location of the file(table) to be created>>'

tblproperties("orc.compress"="SNAPPY")

- Loading table in Amazon S3 Bucket (Bucket should be created, ORC file name should NOT be present already)
- Parquet - Built for complecx data types,All the meatadata written at the end of parquet file.

COMPLEX QUERIES - for statistical analysis

- UDF - User Defined Functions (example : Date objects ,timestamp to date)
- UDAF - aggregate function - bunch of rows as input and one value result
- UDTF - User Defined table generating function - take one row as input and give multiple outputs
- Few examples are : explode,ngrams,sentences,lower,size etc

Code Description

File Name : yelp_hive_queries.hql

File Description : This Hive file contains the table creation and query statements Execute the queries in the same order.

Steps to Run

There are two ways to get into HIVE prompt

- From Windows OS or Linux systems
- In AWS, Create (Hadoop,Hive) Cluster with capacity of m4.xLarge on EMR with 3 instances.
- In case of windows , use Putty to connect to Command line Interface and use WinScp to transfer files from Local Machine to HDFS.
- Uses the public IP allocated to connect using Putty and WinScp.

Hadoop Project-Analysis of Yelp Dataset using Hadoop Hive

The goal of this hadoop project is to apply some data engineering principles to Yelp Dataset in the areas of processing, storage, and retrieval.

□□□□□□□□ □□□□□□□□□:
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.

□□□□ □□□□□□□□□:
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.

□□□□ □□□□□:
➔Language: HQL
➔Services: AWS EMR, Hive, HDFS, AWS S3

□□□ □□□:
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.

□□□□:
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.

□□□ □□□□□□□□□:
● 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.