Assignment 3.3

Explain the below concepts with an example in brief.

• Hive Data Definitions

Hive DDL is similar or can be said a subset of SQL DDL statements that define the structure of the data structure by CREATE, DROP and ALTER of databases, tables, views, partitions and buckets.

a. CREATE Database
b. CREATE Table
c. CREATE View
d. CREATE Function
e. DROP Database
f. DROP Table
g. DROP View
h. DROP Function
i. ALTER Table
j. ALTER Partition

Examples:

CREATE DATABASE my_db;

This will create a database with name my_db. If the DB already exists it will throw error. To avid this error IF NOT EXISTS can be used with CREATE statement.

DROP TABLE my_db;

This will drop the table named my_db.

• Hive Data Manipulations

Like SQL DML, Hive DML provides means to manipulate data by inserting data, sorting, merging tables , updating data etc. It is used to manipulate and transform the data.

a. LOAD – Used for loading data
b. INSERT – Used for inserting data to hive tables, directories also SQL tables.
c. UPDATE – Used to update a particular data present in table.
d. DELETE – Used to delet a particular data from table.
e. MERGE – Merges table.
f. EXPORT – Export data.
g. IMPORT - Import Data.

Examples:

INSERT INTO TABLE my_table VALUES (‘Adam’,’32’),(‘John’, 28);

This will insert data uinto table named my_table.

UPDATE my_table SET age=33 WHERE name=’John’;

This will change the value of the existing data based on condition.

• HiveQL Manipulations

HQL is a SQL like language used in Hive. It reduces the complexity of MapReduce by making it similar to SQL. It reuses familiar concepts of RDBMS.

Hive supports different file formats like TEXTFILE, SEQUENCEFILE, ORC, RC etc.

Along with SQL like queries HQL introduces concepts of optimizing data storage, manipulation of data based on table definition etc.

Examples:

CREATE TABLE IF NOT EXISTS my_tbl (name String, age int)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ‘,’
STORED AS ORC;

Here it creates table like in SQL but additionally it also tells the type of data. So that when the data is stored it the table’s location, it can read the data based on this criteria. It also defines the storing format (ORC) which makes it optimized.