Assignment No. 3

Title:

Write an application using HiveQL for flight information system which will include-

a. Creating, Dropping, and altering Database tables.

b. Creating an external Hive table.

c. Load table with data,insert new values and field in the table, Join tables with Hive.

d. Create index on Flight InformationTable.

e. Find the average departure delay per day in 2008.

Objectives: 1) To describe the basics of Hive.

2) Explain the components of the Hadoop ecosystem.

Aim: To execute a Hive that will perform CRUD operation on Flight Table.

Theory:-

Hive – Introduction

Hive is defined as a data warehouse system for Hadoop that facilitates ad-hoc queries and theanalysis of large datasets stored in Hadoop.

Following are the facts related to Hive:

- It provides a SQL-like language called **HiveQL(HQL).** Due to its SQL-like interface, Hiveis a popular choice for Hadoop analytics.
- It provides massive scale-out and faults tolerance capabilities for data storage andprocessing of commodity hardware.
- Relying on MapReduce for execution, Hive is batch-oriented and has high latency for queryexecution.

Hive – Characteristics

- Hive is a system for managing and querying unstructured data into a structured format.
- It uses the concept of MapReduce for the execution of its scripts and the Hadoop DistributedFile System or HDFS for storage and retrieval of data.

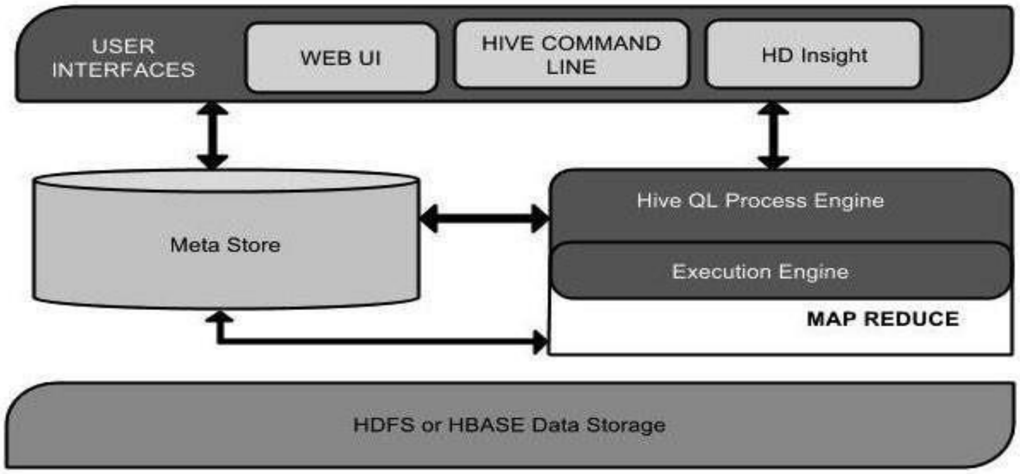
Following are the **key principles** underlying Hive-

- Hive commands are similar to that of SQL. SQL is a data warehousing tool that is similar to Hive.

- Hive contains extensive, pluggable MapReduce scripts in the language of your choice.These scripts include rich, user-defined data types and user-defined functions.
- Hive has an extensible framework to support different files and data formats.
- Performance is better in Hive since Hive engine uses the best-inbuilt script to reduce theexecution time, thus enabling high output in less time.

Architecture of Hive

The following component diagram depicts the architecture of Hive:



Unit Name	Operation
User Interface	Hive is a data warehouse infrastructure software that can create interaction between user and HDFS. The user interfaces that Hive supports are Hive Web UI, Hive command line, and Hive HD Insight.
Meta Store	Hive chooses respective database servers to store the schema or Metadata of tables, databases, columns in a table, their data types, and HDFS mapping.
HiveQL Process Engine	HiveQL is similar to SQL for querying on schema info on the Metastore. Instead of writing MapReduce program in Java, we can write a query for MapReduce job and process it.
Execution Engine	The conjunction part of HiveQL process Engine and MapReduce is Hive Execution Engine. Execution engine

	processes the query and generates results as same as MapReduce results.
HDFS or HBASE	Hadoop distributed file system or HBASE are the data storage techniques to store data into file system.

Listing 13-1: Installing Apache Hadoop and Hive

```
$ mkdirhadoop; cp hadoop-1.2.1.tar.gz hadoop; cd hadoop
$ gunzip hadoop-1.2.1.tar.gz
$ tar xvf *.tar
$ mkdir hive; cp hive-0.11.0.tar.gz hive; cd hive
$ gunzip hive-0.11.0.tar.gz
$ tar xvf *.tar
```

Listing 13-2: Setting Up Apache Hive Environment Variables in .bashrc

```
export HADOOP_HOME=/home/user/Hive/hadoop/hadoop-1.2.1 export
JAVA_HOME=/opt/jdkexport HIVE_HOME=/home/user/Hive/hive-0.11.0 export
PATH=$HADOOP_HOME/bin:$HIVE_HOME/bin:
$JAVA_HOME/bin:$PATH
```

Listing 13-3: Setting Up the hive-site.xml File

```
$ cd $HIVE_HOME/conf
$ cp hive-default.xml.templateto hive-site.xml
```

Working with Hive Data Types

Listing 13-7 goes to the trouble of creating a table that uses all Hive-supported data types and theamount of memory required.

Commands for Hive

Create Database Statement In Hive

```
hive> CREATE DATABASE userdb;
```

Create Table Statement

```
hive> CREATE TABLE IF NOT EXISTS employee ( eidint, name String,

salary String, destination String)COMMENT „Employee details‟ROW

        FORMAT DELIMITED FIELDS TERMINATED

BY „\t‟LINESTERMINATED BY „\n‟STORED AS

TEXTFILE;
```

Alter Table Statement

```
hive> ALTER TABLE employee RENAME TO emp;
```

Drop Table Statement

```
hive>DROP TABLE IF EXISTS employee;
```

1) **Creating, Dropping, and altering Database tables**

hbase(main):001:0> create 'flight','finfo','fsch'

2) **Load table with data, insert new values and field in the table, Join tables with Hivehbase(main):002:0> put 'flight',1,'finfo:dest','mumbai'**

3) **Create index on Flight information Table**

```
hive>CREATE INDEX ine ON TABLE FLIGHT(source) AS
'org.apache.hadoop.hive.ql.index.compact.CompactIndexHandler' WITH
DEFERRED REBUILD;
```

4) **Find the average departure delay per day in 2008.**

```
hive>select avg(delay) from flight where year = 2008;
```

Conclusion: Thus, we have learnt HiveQL.