Assignment No: 3

Aim: Study of Hive

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

Prerequisites:

- Ensure that Hadoop is installed, configured and is running.
- Single Node Setup.
- Hive installed and working propery
- Hbase installed and working propery

Theory:

Hive : Hive is a data warehouse infrastructure tool used to process data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy.

Steps :

[1] Start Hadoop
[2] Start Hive
[3] Create a database
Example :
Hive>CREATE DATABASE ourfirstdatabase;

Hive> USE ourfirstdatabase;

[4] Crete a table

```
hive >CREATE TABLE our_first_table
                                    (
                                        FirstName STRING,
                                        LastName STRING,
                                        EmployeeId INT
                                    );
```

Examples:

hive> DROP DATABASE ourfirstdatabase CASCADE;

3) Download Flight data set of 2007 & 2008 from :
http://stat-computing.org/dataexpo/2009/the-data.html

Dataset description: Different fields in the flight data set are

	Name	Description
1	Year	1987-2008
2	Month	1-12
3	DayofMonth	1-31
4	DayOfWeek	1 (Monday) - 7 (Sunday)
5	DepTime	actual departure time (local, hhmm)
6	CRSDepTime	scheduled departure time (local, hhmm)
7	ArrTime	actual arrival time (local, hhmm)
8	CRSArrTime	scheduled arrival time (local, hhmm)
9	UniqueCarrier	unique carrier code
10	FlightNum	flight number
11	TailNum	plane tail number
12	ActualElapsedTime	in minutes
13	CRSElapsedTime	in minutes
14	AirTime	in minutes
15	ArrDelay	arrival delay, in minutes
16	DepDelay	departure delay, in minutes
17	Origin	origin IATA airport code
18	Dest	destination IATA airport code
19	Distance	in miles
20	TaxiIn	taxi in time, in minutes
21	TaxiOut	taxi out time in minutes
22	Cancelled	was the flight cancelled?
23	CancellationCode	reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
24	Diverted	1 = yes, 0 = no
25	CarrierDelay	in minutes
26	WeatherDelay	in minutes
27	NASDelay	in minutes
28	SecurityDelay	in minutes
29	LateAircraftDelay	in minutes

Create a table:

```
CREATE TABLE IF NOT EXISTS FlightInfo2007
  (
        Year SMALLINT, Month TINYINT, DayofMonth TINYINT,
        DayOfWeek TINYINT,
        DepTime SMALLINT, CRSDepTime SMALLINT, ArrTime SMALLINT,CRSArrTime SMALLINT,
          UniqueCarrier STRING, FlightNum STRING, TailNum STRING,
          ActualElapsedTime SMALLINT, CRSElapsedTime SMALLINT,
          AirTime SMALLINT, ArrDelay SMALLINT, DepDelay SMALLINT,
          Origin STRING, Dest STRING,Distance INT,
          TaxiIn SMALLINT, TaxiOut SMALLINT, Cancelled SMALLINT,
         CancellationCode STRING, Diverted SMALLINT,
          CarrierDelay SMALLINT, WeatherDelay SMALLINT,
         NASDelay SMALLINT, SecurityDelay SMALLINT,
         LateAircraftDelay
         SMALLINT)
                  COMMENT 'Flight InfoTable'
                  ROW FORMAT DELIMITED
                  FIELDS TERMINATED BY ','
                  LINES TERMINATED BY '\n'
```

Load Data into table:

```
hive> load data local inpath '/home/hduser/Desktop/2007.csv' into table FlightInfo2007;

hive> CREATE TABLE IF NOT EXISTS FlightInfo2008 LIKE FlightInfo2007;
hive> load data local inpath '/home/hduser/Desktop/2008.csv' into table FlightInfo2008;

hive> CREATE TABLE IF NOT EXISTS myFlightInfo (
                  Year SMALLINT, DontQueryMonth TINYINT, DayofMonth
                  TINYINT, DayOfWeek TINYINT, DepTime SMALLINT, ArrTime SMALLINT,
                  UniqueCarrier STRING, FlightNum STRING,
                  AirTime SMALLINT, ArrDelay SMALLINT, DepDelay SMALLINT,
                  Origin STRING, Dest STRING, Cancelled SMALLINT,
                  CancellationCode STRING)
                  COMMENT 'Flight InfoTable'
                  PARTITIONED BY(Month TINYINT)
                  ROW FORMAT DELIMITED
                  FIELDS TERMINATED BY ','
                  LINES TERMINATED BY '\n' ;
```

```
hive> CREATE TABLE myflightinfo2007 AS SELECT Year, Month, DepTime, ArrTime,
    FlightNum, Origin, Dest FROM FlightInfo2007 WHERE (Month = 7 AND DayofMonth = 3)
    AND (Origin='JFK' AND Dest='ORD');

hive>SELECT * FROM myFlightInfo2007;


hive> CREATE TABLE myFlightInfo2008 AS SELECT Year, Month, DepTime, ArrTime,
    FlightNum, Origin, Dest FROM FlightInfo2008 WHERE (Month = 7 AND DayofMonth = 3)
    AND (Origin='JFK' AND Dest='ORD');
hive> SELECT * FROM myFlightInfo2008;
```

JOIN

```
Hive>SELECT m8.Year, m8.Month, m8.FlightNum, m8.Origin, m8.Dest, m7.Year, m7.Month,
    m7.FlightNum, m7.Origin, m7.Dest FROM myFlightinfo2008 m8 JOIN myFlightinfo2007
    m7 ON m8.FlightNum=m7.FlightNum;

hive> SELECT m8.FlightNum,m8.Origin,m8.Dest,m7.FlightNum,m7.Origin,m7.Dest FROM
    myFlightinfo2008 m8 FULL OUTER JOIN myFlightinfo2007 m7 ON
    m8.FlightNum=m7.FlightNum;

hive>SELECT
    m8.Year,m8.Month,m8.FlightNum,m8.Origin,m8.Dest,m7.Year,m7.Month,m7.FlightNum,
    m7.Origin,m7.Dest FROM myFlightinfo2008 m8 LEFT OUTER JOIN myFlightinfo2007
    m7 ON m8.FlightNum=m7.FlightNum;



hive> CREATE INDEX f08_index ON TABLE flightinfo2008 (Origin) AS
  > 'COMPACT' WITH DEFERRED REBUILD;

hive> ALTER INDEX f08_index ON flightinfo2008 REBUILD;

hive>SHOW INDEXES ON FlightInfo2008;


hive> SELECT Origin, COUNT(1) FROM flightinfo2008 WHERE Origin = 'SYR' GROUP BY
    Origin;

hive> DESCRIBE default__flightinfo2008_f08_index__;

hive> CREATE VIEW avgdepdelay AS SELECT DayOfWeek, AVG(DepDelay) FROM
   FlightInfo2008 GROUP BY DayOfWeek;

hive> SELECT * FROM avgdepdelay;

3    8.289761053658728
```

6 8.645680904903614
1 10.269990244459473
4 9.772897177836702
7 11.568973392595312
2 8.97689712068735
5 12.158036387869656

Day 5 under the results in Step (B) — had the highest number of delays.

Conclusion : We have studied hive for big data analysis