**AIRLINE ON-TIME DATA ANALYSIS (HIVE ASSIGNMENT)**

**Step 1.**

1. hdfs dfs -mkdir /user/hadoop/AirlinesOntimeAnalysis

hdfs dfs -copyFromLocal airports.csv /user/hadoop/AirlinesOntimeAnalysis

hdfs dfs -copyFromLocal carriers.csv /user/hadoop/AirlinesOntimeAnalysis

hdfs dfs -copyFromLocal plane-data.csv /user/hadoop/AirlinesOntimeAnalysis

hdfs dfs -copyFromLocal flights.csv /user/hadoop/AirlinesOntimeAnalysis

1. find /user/hadoop -name "\*.bz2" -exec bzip2 -d {} \;
2. hadoop fs -mkdir -p /user/$USER/rawdata/airline/flights

hadoop fs -mkdir -p /user/$USER/rawdata/airline/planeinfo

hadoop fs -mkdir -p /user/$USER/rawdata/airline/airport

hadoop fs -mkdir -p /user/$USER/rawdata/airline/carriers

1. hdfs dfs -copyFromLocal airports.csv /rawdata/airline/airport

hdfs dfs -copyFromLocal carriers.csv /rawdata/airline/carriers

hdfs dfs -copyFromLocal plane-data.csv /rawdata/airline/planeinfo

hdfs dfs -copyFromLocal flights.csv /rawdata/airline/flights

**Step 2.**

Create database IF NOT EXISTS airline\_db;

Use airline\_db;

Create external table flights\_stg (

Year INT,

Month INT,

DayofMonth INT,

DayofWeek INT,

DepTime INT

CRSDepTime INT,

ArrTime INT,

CRSArrTime INT, scheduled arrival time (local, hhmm)

UniqueCarrier STRING,

FlightNum STRING,

TailNum STRING,

ActualElapsedTime INT,

CRSElapsedTime INT,

* Remaining columns

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LOCATION '/user/hive/warehouse/airline\_db.db/flights/';

Create external table airports\_stg (

IATA STRING,

Name STRING,

City STRING,

State STRING,

Country STRING,

Lat DOUBLE

Long DOUBLE

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LOCATION '/user/hive/warehouse/airline\_db.db/airports/';

Create external table carriers\_stg (

Code STRING,

Description STRING,

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LOCATION '/user/hive/warehouse/airline\_db.db/carriers/';

Create external table plane\_info\_stg (

Type STRING,

manufacturer STRING,

issue date STRING,

model STRING,

status STRING,

aircraft\_type STRING,

engine\_type STRING,

year INT

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LOCATION '/user/hive/warehouse/airline\_db.db/plane\_info/';

Check if the tables have data

**Solution**.

Select \* from flights\_stg;

Select \* from airport\_stg;

Select \* from carriers\_stg;

Select \* from olane\_info\_stg;

**What is the average arrival delay and average departure delay in each month of the year 2004?**

**Solution**:

SELECT month,

AVG(ArrDelay) AS avg\_arrival\_delay,

AVG(DepDelay) AS avg\_departure\_delay

FROM flights\_stg

WHERE YEAR= 2004

GROUP BY month

ORDER BY month;

**Step 3.**

Create external table flights\_pq (

Year INT,

Month INT,

DayofMonth INT,

DayofWeek INT,

DepTime INT

CRSDepTime INT,

ArrTime INT,

CRSArrTime INT, scheduled arrival time (local, hhmm)

UniqueCarrier STRING,

FlightNum STRING,

TailNum STRING,

ActualElapsedTime INT,

CRSElapsedTime INT,

* Remaining columns

)

STORED AS PARQUET

LOCATION '/user/hive/output/airline/flights\_pq/'

TBLPROPERTIES ('parquet.compress'='SNAPPY');

Create external table airports\_pq (

IATA STRING,

Name STRING,

City STRING,

State STRING,

Country STRING,

Lat DOUBLE

Long DOUBLE

)

STORED AS PARQUET

LOCATION '/user/hive/output/airline/airport\_pq/'

TBLPROPERTIES ('parquet.compress'='SNAPPY');

Create external table carriers\_pq (

Code STRING,

Description STRING,

)

STORED AS PARQUET

LOCATION '/user/hive/output/airline/carriers\_parquet/'

TBLPROPERTIES ('parquet.compress'='SNAPPY');

Create external table plane\_info\_pq (

Type STRING,

manufacturer STRING,

issue date STRING,

model STRING,

status STRING,

aircraft\_type STRING,

engine\_type STRING,

year INT

)

STORED AS PARQUET

LOCATION '/user/hive/output/airline/plane\_info\_pq/'

TBLPROPERTIES ('parquet.compress'='SNAPPY');

**Is there a difference between the response times for the queries(2 and 3)**

**Solution**. YES