Create airport table.

hive> Create external table airports\_savu (IATA String, Name String, City String, State String, Country String, Lat Decimal (10,6), Long Decimal (10,6) )ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' location '/user/mavricbdhoct07/airline/airports' TBLPROPERTIES ("skip.header.line.count"="1");

hive> load data local inpath 'airports.csv' OVERWRITE into table airports\_savu;

A close-up of a computer screen

Description automatically generated

Create Table carrier

hive> Create external table carrier\_savu (code string , description string )ROW FORMAT DELIMITED FIELDS TERMINATEDBY ',' location '/user/mavricbdhoct07/airline/carriers' TBLPROPERTIES ("skip.header.line.count"="1");

hive> load data local inpath 'carriers.csv' OVERWRITE into table carrier\_savu;

A screenshot of a computer

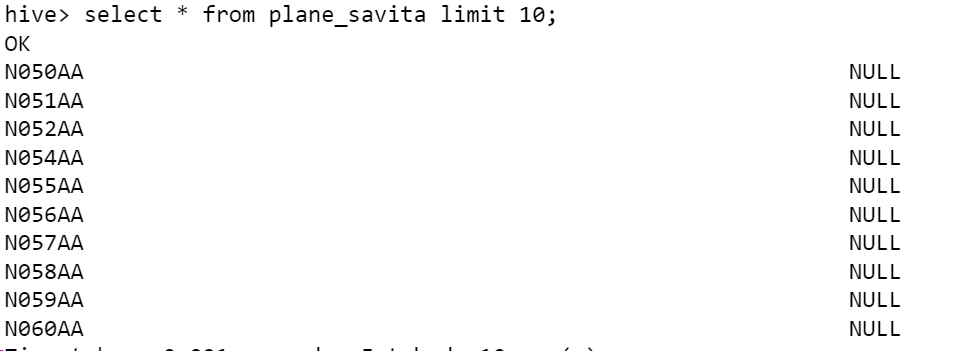
Description automatically generated

Create Plane Table

hive> Create external table plane\_savita (tailnum string,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/mavricbdhoct07/airline/planeinfo'

TBLPROPERTIES ("skip.header.line.count"="1");

hive> load data local inpath 'plane-data.csv' OVERWRITE into table plane\_savita;

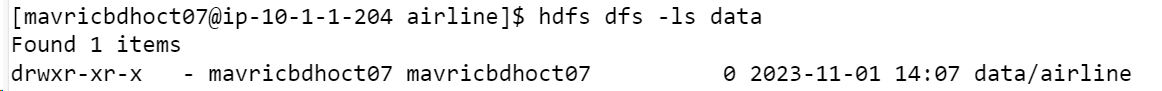


Create flights table

hive> create external table flights\_savu (Year Int, Month Int, DayofMonth Int, DayOfWeek Int,DepTime Int,CRSDepTime Int,ArrTime Int, CRSArrTime Int, UniqueCarrier String, FlightNum String, TailNum String, ActualElapsedTime Int, CRSElapsedTime Int, AirTime Int, ArrDelay Int, DepDelay Int, Origin String, Dest String, Distance Int, TaxiIn String, TaxiOut String, Cancelled String, CancellationCode String, Diverted String, CarrierDelay Int, WeatherDelay Int, NASDelay Int, SecurityDelay Int, LateAircraftDelay Int) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' location '/user/mavricbdhoct07/airline/flights' TBLPROPERTIES ("skip.header.line.count"="1");

hive> load data local inpath '.csv' OVERWRITE into table flights\_savu;

[mavricbdhoct07@ip-10-1-1-204 airline]$ hdfs dfs -mkdir data/airline



[mavricbdhoct07@ip-10-1-1-204 airline]$ hdfs dfs -mkdir data/airline/flights

[mavricbdhoct07@ip-10-1-1-204 airline]$ hdfs dfs -mkdir data/airline/airports

[mavricbdhoct07@ip-10-1-1-204 airline]$ hdfs dfs -mkdir data/airline/planeinfo

[mavricbdhoct07@ip-10-1-1-204 airline]$ hdfs dfs -mkdir data/airline/carrie

1.Data Setup in HDFS

Step 1: Copy the provided csv files in a directory named AirlinesOntimeAnalysis

[mavricbdhoct07@ip-10-1-1-204 ~]$ hdfs dfs -mkdir AirlinesOntimeAnalysis

Step 2: Unzip the contents for the bz2 files.

[mavricbdhoct07@ip-10-1-1-204 ~]$ cd airline [mavricbdhoct07@ip-10-1-1-204 airline]$ bzip2 -d 2003.csv.bz2

[mavricbdhoct07@ip-10-1-1-204 airline]$ bzip2 -d 2005.csv.bz2

[mavricbdhoct07@ip-10-1-1-204 airline]$ bzip2 -d 2004.csv.bz2

Step 3: Create the directories in hdfs where your data will reside

Step 4: copy the respective files to the above created hdfs

structureA white background with black text

Description automatically generated

2.Create tables using Hive

hive> !hdfs dfs -ls /user/hive/warehouse/airlinedb.db;

A close-up of a document

Description automatically generated

3. Improving the performance of Hive tables - Parquet

Format

[mavricbdhoct07@ip-10-1-1-204 ~]$ ls airline

2003.csv 2004.csv 2005.csv airports.csv carriers.csv plane-data.csv

[mavricbdhoct07@ip-10-1-1-204 ~]$ hdfs dfs -put airline/2003.csv AirlinesOntimeAnalysis

[mavricbdhoct07@ip-10-1-1-204 ~]$ hdfs dfs -put airline/2004.csv AirlinesOntimeAnalysis

[mavricbdhoct07@ip-10-1-1-204 ~]$ hdfs dfs -put airline/2005.csv AirlinesOntimeAnalysis

[mavricbdhoct07@ip-10-1-1-204 ~]$ hdfs dfs -put airline/airports.csv AirlinesOntimeAnalysis

[mavricbdhoct07@ip-10-1-1-204 ~]$ hdfs dfs -put airline/carriers.csv AirlinesOntimeAnalysis

[mavricbdhoct07@ip-10-1-1-204 ~]$ hdfs dfs -put airline/plane-data.csv AirlinesOntimeAnalysis

[mavricbdhoct07@ip-10-1-1-204 ~]$ hdfs dfs -ls AirlinesOntimeAnalysis

A black text on a white background

Description automatically generated

Load the data

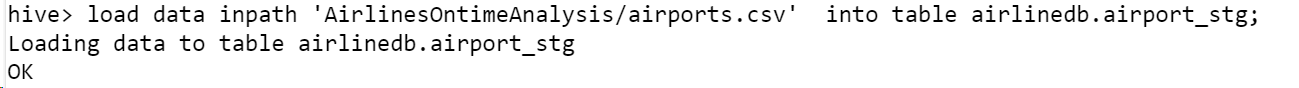
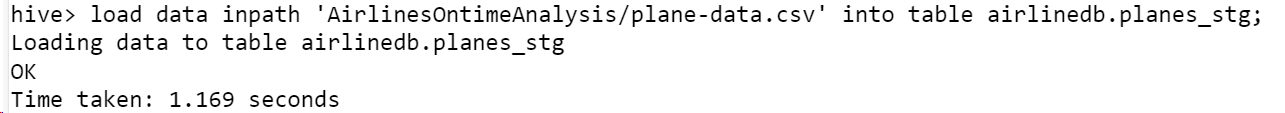
A close-up of a document

Description automatically generated

hive> load data inpath 'AirlinesOntimeAnalysis/airports.csv' into table airlinedb.airport\_stg;

hive> load data inpath 'AirlinesOntimeAnalysis/carriers.csv' into table airlinedb.carriers\_stg;

hive> load data inpath 'AirlinesOntimeAnalysis/plane\_data.csv' into table airlinedb.planes\_stg;

A computer screen with text

Description automatically generated

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

hive> select month , avg(arrdelay) as arrival\_delay , avg(depdelay) as departure\_delay from airlinedb.flight\_stg where year=2004 group by month;

A white background with black text

Description automatically generated

3.Improving the performance of Hive tables - Parquet format

a.Create the parquet table as external tables for flights from the above staging table.

hive> create external table flights\_pq (Year Int, Month Int, DayofMonth Int, DayOfWeek Int,DepTime Int,CRSDepTime Int,ArrTime Int, CRSArrTi

me Int, UniqueCarrier String, FlightNum String, TailNum String, ActualElapsedTime Int, CRSElapsedTime Int, AirTime Int, ArrDelay Int, DepD

elay Int, Origin String, Dest String, Distance Int, TaxiIn String, TaxiOut String, Cancelled String, CancellationCode String, Diverted Stri

ng, CarrierDelay Int, WeatherDelay Int, NASDelay Int, SecurityDelay Int, LateAircraftDelay Int) STORED AS PARQUET location '/user/mavricbd

hoct07/output/airline/pq\_flights';

**b.Name the new table with \_pq as a suffix: Ex : flights\_pq**

insert the values to flights\_pq

hive> insert overwrite table flights\_pq select \* from flight\_stg;

A close up of a text

Description automatically generated

**Run the same query as above to check if data is loaded properly**

hive> select \* from flights\_pq limit 10;

OK

A close-up of a number

Description automatically generated

hive> select Month,avg(ArrDelay) as arrival\_delay, avg(DepDelay) as depature\_delay from flights\_pq where year=2004 group by month;

A white screen with black text

Description automatically generated

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

Yes, both 2 and 3 query response time is difference.