**Data Setup in HDFS:**

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

A close-up of a document

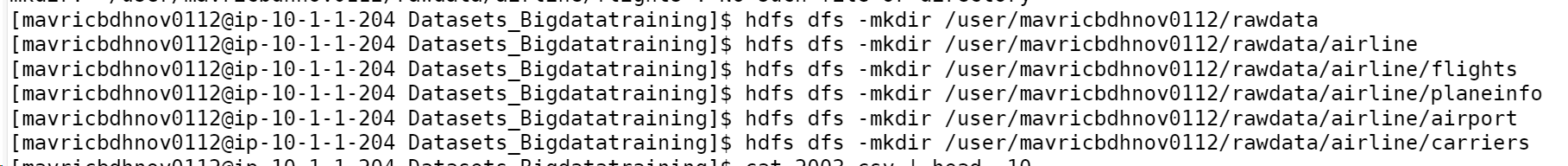
Description automatically generated

2.Unzip the contents for the bz2 files:

A computer screen shot of a computer program

Description automatically generated

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



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

1. **Create tables using Hive**

Create the required tables as External (flights, airports, carriers, plane\_info) in the staging area. The schema of each of the tables is provided above. Also point these tables to the hdfs locations already created above Note: Name every table name at the end with \_stg. Create all the tables in a different database in Hive(airline\_db

**Create database airline\_db;**

Create external table airports\_stg(

IATA String, Name String,City String,State String,Country String,Lat Decimal,Long Decimal)

Row format delimited fields terminated by “,”

Lines terminated by “\n”

Stored as textfile

Location ‘/user/mavricbdhnov0112/airline/airport’

Tblproperties(“skip.header.line.count”=”1”);

A computer screen shot of a computer code

Description automatically generated

Create external table carrier\_stg(

code String, description String)

Row format delimited fields terminated by ","

Lines terminated by "\n"

Stored as textfile

Location '/user/mavricbdhnov0112/airline/carriers'

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

A computer code with text

Description automatically generated

Create external table planeinfo\_stg(

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 ","

Lines terminated by "\n"

Stored as textfile

Location '/user/mavricbdhnov0112/airline/planeinfo'

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

A white background with black text

Description automatically generated

Create external table flight\_stg(

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 ","

Lines terminated by "\n"

Stored as textfile

Location '/user/mavricbdhnov0112/airline/flights'

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

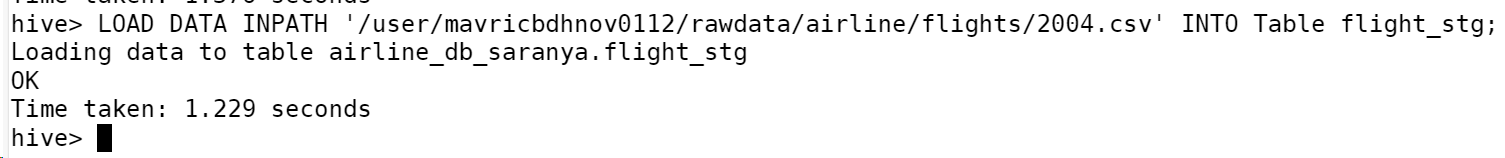
A screenshot of a computer

Description automatically generated

Create tables for flights , plane\_info, carriers and airports in the staging area.

A close-up of a computer code

Description automatically generated



Check if the tables have data.

A close-up of a text

Description automatically generated

A close-up of a number

Description automatically generated

A white background with black text

Description automatically generated

A close up of a document

Description automatically generated

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

**Select avg(arrdelay) as avg\_arrival\_delay ,avg(depdelay) as avg\_departure\_delay from flight\_stg group by month;**

A screenshot of a computer program

Description automatically generated

1. **Improving the performance of Hive tables** - Parquet format

Create the parquet table as external tables for flights from the above staging table. Name the new table with \_pq as a suffix: Ex : flights\_pq Point all the table location output/airline/pq\_flights

Create external table flight\_stg\_pq(

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 ","

Lines terminated by "\n"

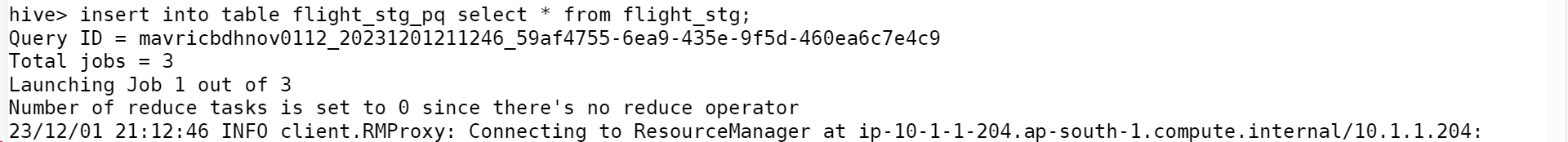
Stored as parquet

Location '/user/mavricbdhnov0112/output/airline/pq\_flights'

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

A close-up of a document

Description automatically generated



Run the same query as above to check if data is loaded properly Is there a difference between the response times for the queries(2 and 3)

There is difference of 7.38 seconds for query 2 and 3.

A white text with black text

Description automatically generated with medium confidence