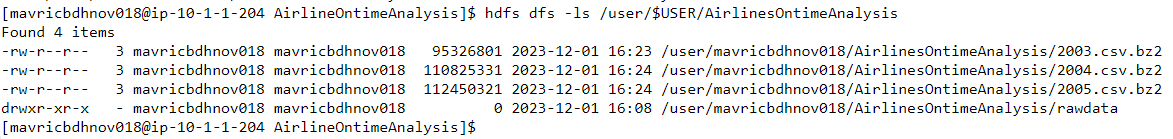
**Hive Assignment 3**

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



**Step 2:** Unzip the contents for the bz2 files.

A screenshot of a computer

Description automatically generated

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

A close-up of a text

Description automatically generated

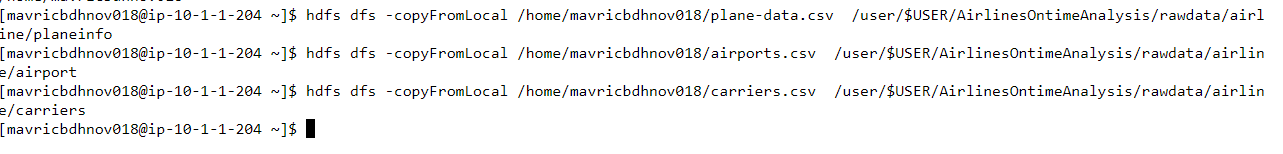
**Step 4: copy the respective files to the above created hdfs structure**

>>hdfs dfs -put AirlinesOntimeAnalysis/200\*.csv rawdata/airline/flights

>>hdfs dfs -put AirlinesOntimeAnalysis/plane\_data.csv rawdata/airline/planeinfo

>>hdfs dfs -put AirlinesOntimeAnalysis/airports.csv rawdata/airline/airport

>>hdfs dfs -put AirlinesOntimeAnalysis/carriers.csv rawdata/airline/carriers



**2. Create tables using Hive**

**Create all the tables in a different database in Hive(airline\_db)**

>>CREATE DATABASE airlines\_db;

A screenshot of a computer

Description automatically generated

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

**Hive QL:**

hive> CREATE EXTERNAL TABLE carriers\_stg (Code string,Description string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' l

ocation '/user/$USER/AirlinesOntimeAnalysis/rawdata/airline/carriers';

hive> CREATE EXTERNAL TABLE airport\_stg (iata string,airport string,city string,state string,country string,lat decimal,l

ong decimal) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' location '/user/$USER /AirlinesOntimeAnalysis/rawdata/airline/airport';

hive> CREATE EXTERNAL TABLE plane\_info\_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 ',' location '

/user/$USER/AirlinesOntimeAnalysis/rawdata/airline/planeinfo';

hive> CREATE EXTERNAL TABLE flights\_stg (Year int,Month int,DayofMonth int,DayOfWeek int,DepTime int,CRSDepTime int,ArrTi

me int,CRSArrTime int,UniqueCarrier string,FlightNum string,TailNum string,ActualElapsedTime int,CRSElapsedTime int,AirTi

me int,ArrDelay int,DepDelay int,Origin string,Dest string,Distance int,TaxiIn string,TaxiOut string,Cancelled string,Can

cellationCode string,Diverted string,CarrierDelay int,WeatherDelay int,NASDelay int,SecurityDelay int,LateAircraftDelay i

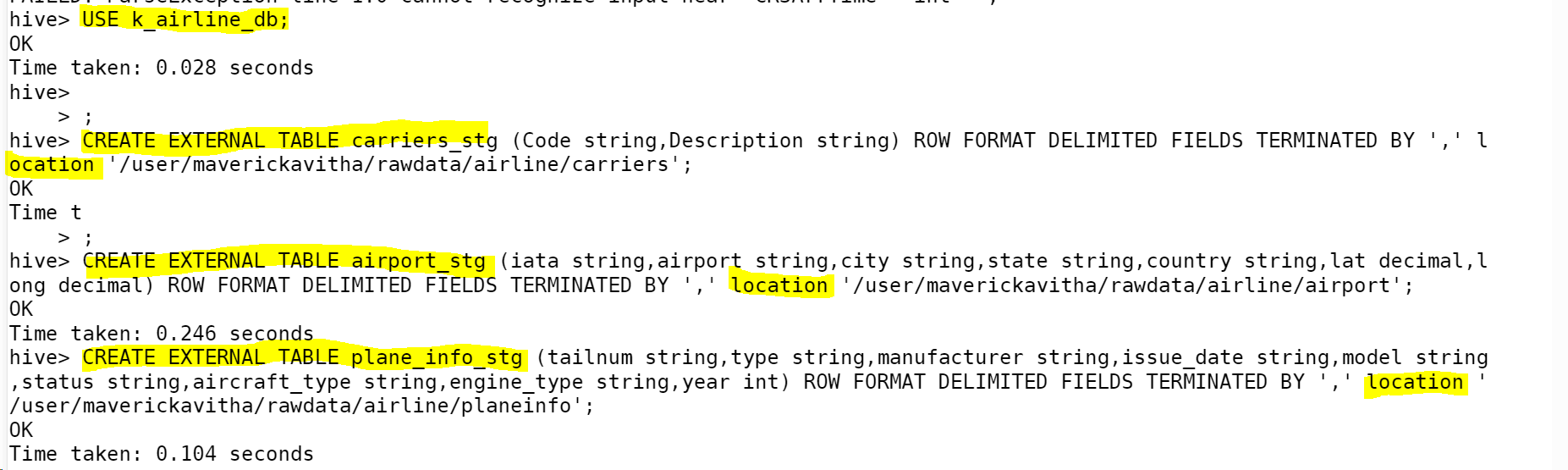
nt) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' location '/user/$USER/AirlinesOntimeAnalysis/rawdata/airline/flights' ;

hive> ALTER TABLE flights\_stg SET TBLPROPERTIES ("skip.header.line.count"="1");

hive> ALTER TABLE plane\_info\_stg SET TBLPROPERTIES ("skip.header.line.count"="1");

hive> ALTER TABLE airport\_stg SET TBLPROPERTIES ("skip.header.line.count"="1");

hive> ALTER TABLE carriers\_stg SET TBLPROPERTIES ("skip.header.line.count"="1");



A yellow text on a white background

Description automatically generated

A screenshot of a computer code

Description automatically generated

**Check if the tables have data.**

select \* from k\_airline\_db.flights\_stg limit 5;

select \* from k\_airline\_db.plane\_info\_stg limit 5;

select \* from k\_airline\_db.airport\_stg limit 5;

select \* from k\_airline\_db.carriers\_stg limit 5;

A screenshot of a computer

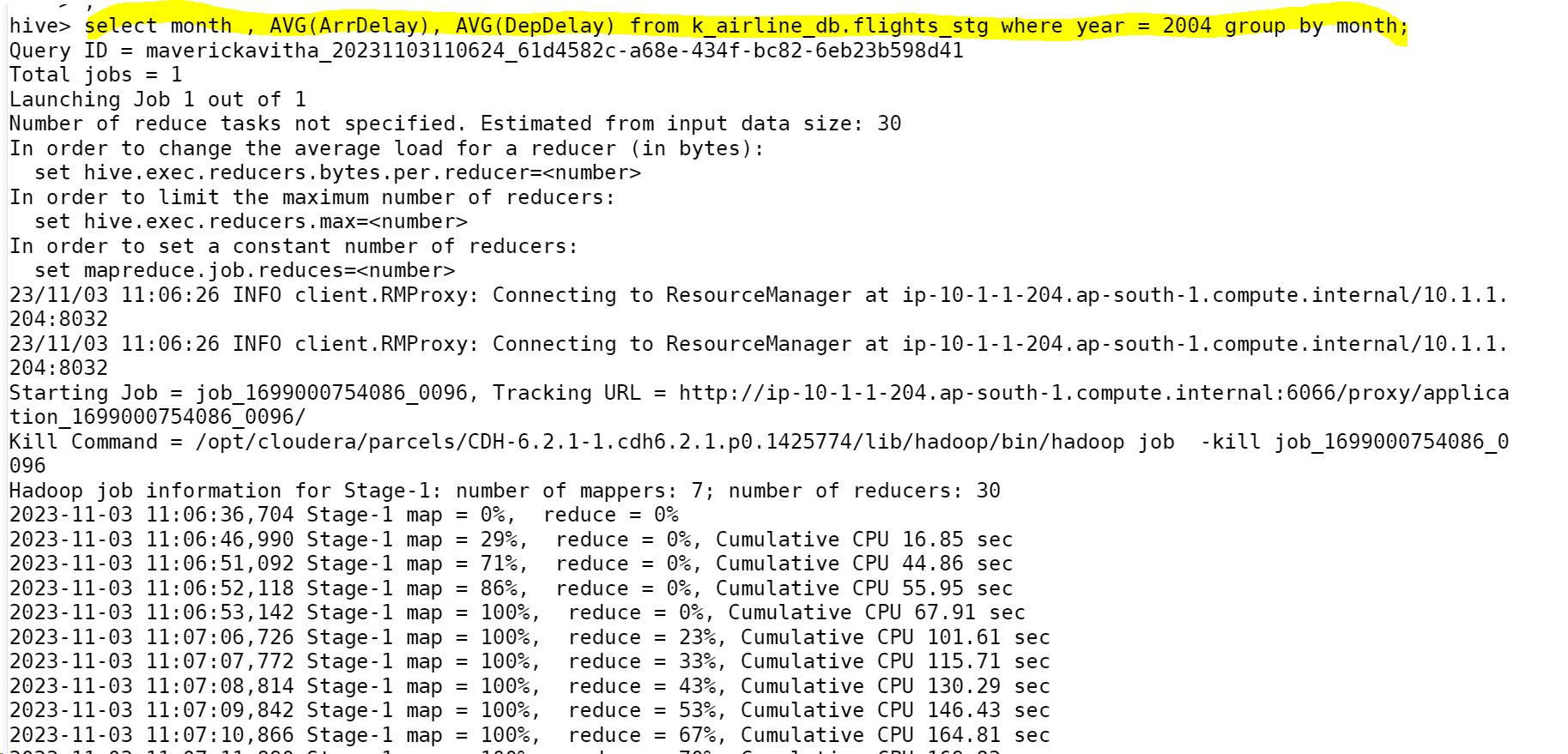
Description automatically generated

A screenshot of a computer

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), AVG(DepDelay) from k\_airline\_db.flights\_stg where year = 2004 group by month;



A screenshot of a computer

Description automatically generated

**3. 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

hive> [maverickavitha@ip-10-1-1-204 ~]$ hdfs dfs -mkdir -p output/airline/pq\_flights

hive> CREATE EXTERNAL TABLE flight\_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,Cance

llationCode string,Diverted string,CarrierDelay int,WeatherDelay int,NASDelay int,SecurityDelay int,LateAircraftDelay int

) STORED AS PARQUET location '/user/maverickavitha/output/airline/pq\_flights';

hive> insert overwrite table k\_airline\_db.flight\_pq select \* from k\_airline\_db.flights\_stg;

A yellow text with black text

Description automatically generated

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

A screenshot of a computer

Description automatically generated

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

A screenshot of a computer

Description automatically generated

**AVG Calculation in PARQUET format:**

A white paper with black text

Description automatically generated

A screenshot of a computer

Description automatically generated

**AVG Calculation in normal STG table:**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated