**Assignment 3 ontime flight performances:**

**Steps completed in HDFS:**

hdfs dfs -mkdir indhuAirlinesOntimeAnalysis

hdfs dfs -put 2003.csv.bz2 indhuAirlinesOntimeAnalysis

hdfs dfs -put 2004.csv.bz2 indhuAirlinesOntimeAnalysis

hdfs dfs -put 2005.csv.bz2 indhuAirlinesOntimeAnalysis

hdfs dfs -put airports.csv indhuAirlinesOntimeAnalysis

hdfs dfs -put carriers.csv indhuAirlinesOntimeAnalysis

hdfs dfs -put plane-data.csv indhuAirlinesOntimeAnalysis

bzip2 -dk 2003.csv.bz2

bzip2 -dk 2004.csv.bz2

bzip2 -dk 2005.csv.bz2

hdfs dfs -put 2003.csv indhuAirlinesOntimeAnalysis

hdfs dfs -put 2004.csv indhuAirlinesOntimeAnalysis

hdfs dfs -put 2005.csv indhuAirlinesOntimeAnalysis

hdfs dfs -mkdir rawdata/airline/airport

hdfs dfs -mkdir rawdata/airline/flights

hdfs dfs -mkdir rawdata/airline/carriers

hdfs dfs -mkdir rawdata/airline/planeinfo

hdfs dfs -cp indhuAirlinesOntimeAnalysis/airports.csv rawdata/airline/airport

hdfs dfs -cp indhuAirlinesOntimeAnalysis/200\* rawdata/airline/flights

hdfs dfs -cp indhuAirlinesOntimeAnalysis/plane-data.csv rawdata/airline/planeinfo

hdfs dfs -cp indhuAirlinesOntimeAnalysis/carriers.csv rawdata/airline/carriers

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

create database indhuairline\_db;

use indhuairline\_db;

create EXTERNAL TABLE airports\_stg (iata string,airport string,city string,state string,country string,lat decimal,long decimal) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' location '/user/mavricbdhoct08/rawdata/airline/airport';

create EXTERNAL TABLE carrier\_stg (code string,description string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' location '/user/mavricbdhoct08/rawdata/airline/carriers';

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 ',' location '/user/mavricbdhoct08/rawdata/airline/planeinfo';

create EXTERNAL TABLE flights\_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 ',' location '/user/mavricbdhoct08/rawdata/airline/flights';

A screenshot of a computer

Description automatically generated

Select \* from airports\_stg limit 10;

A screenshot of a computer

Description automatically generated

Select \* from carrier\_stg limit 10;

A screenshot of a computer

Description automatically generated

Select count(\*) from flights\_stg ;

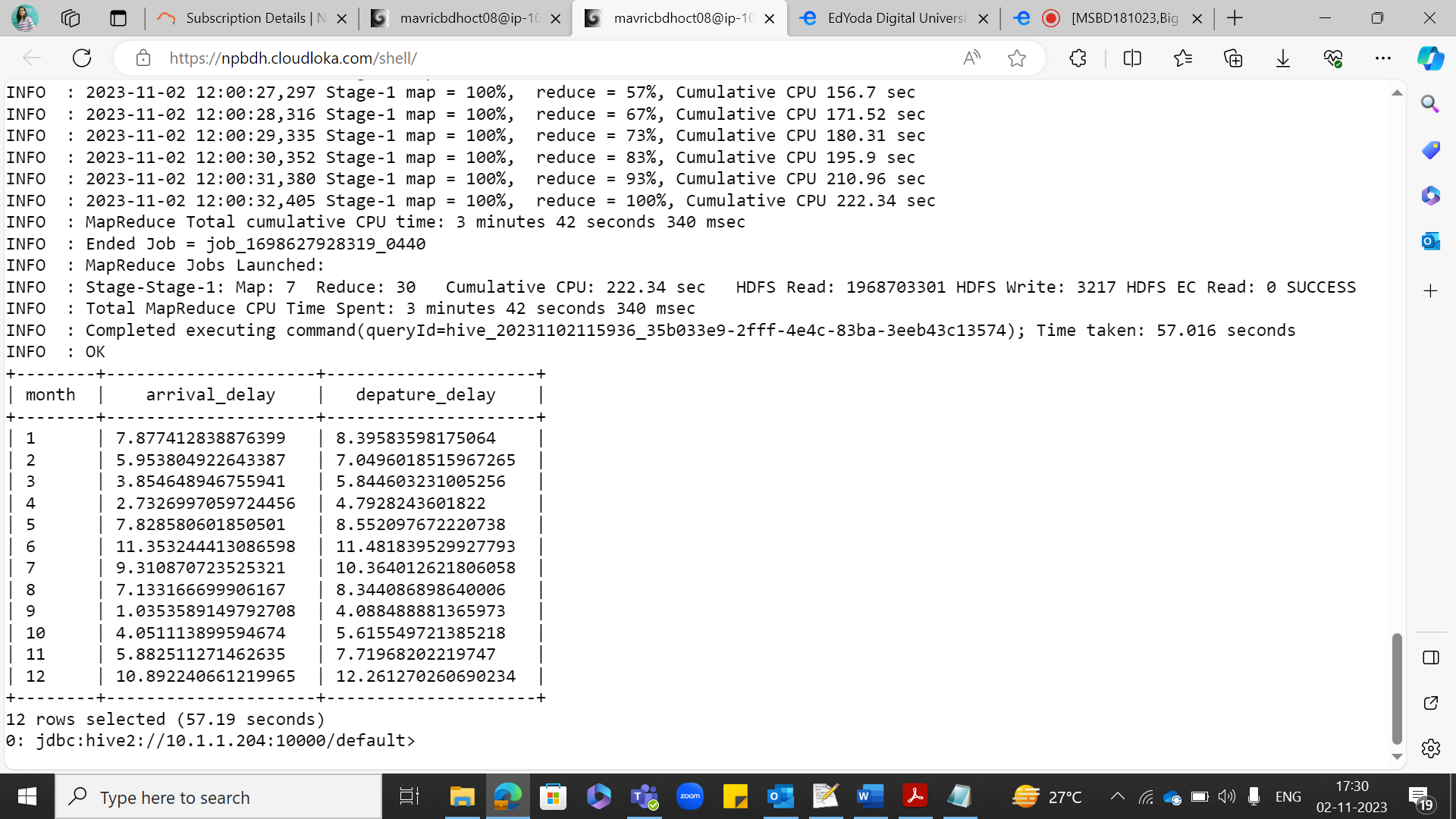
A screenshot of a computer

Description automatically generated

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

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

Time taken: 57.19 seconds



**PARQUET TABLE CREATION:**

create EXTERNAL TABLE flights\_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) STORED AS PARQUET location'/user/mavricbdhoct08/output/airline/pq\_flights';

insert overwrite table flights\_pq select \* from flights\_stg;

A screenshot of a computer

Description automatically generated

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

A screenshot of a computer

Description automatically generated

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

Time taken: 42.403 seconds

A screenshot of a computer

Description automatically generated