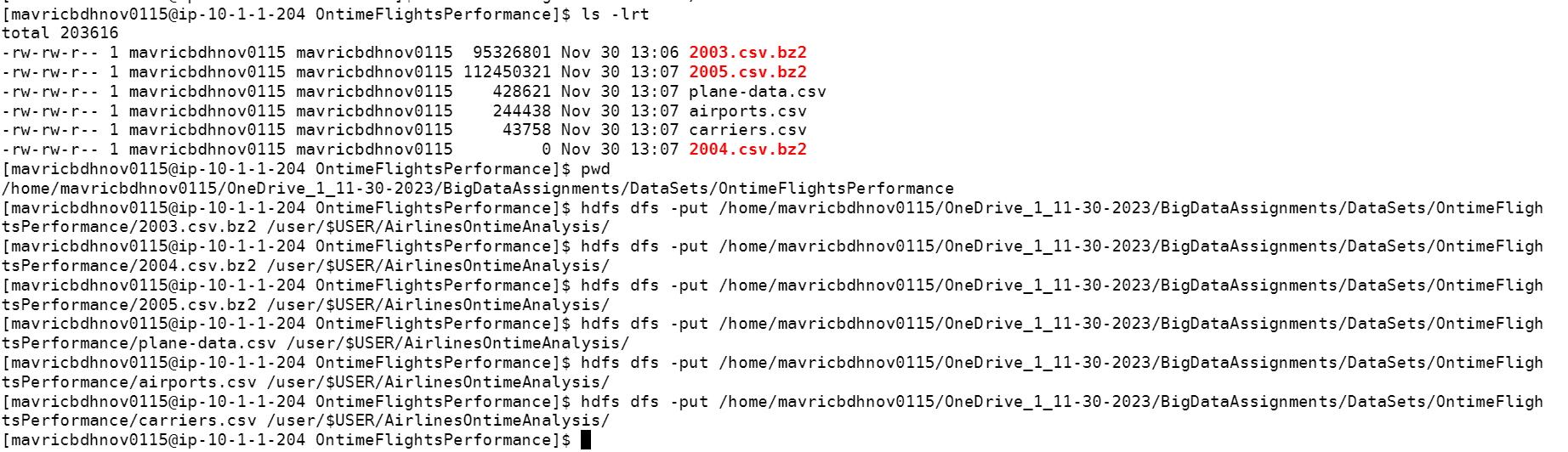
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Problem-1\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

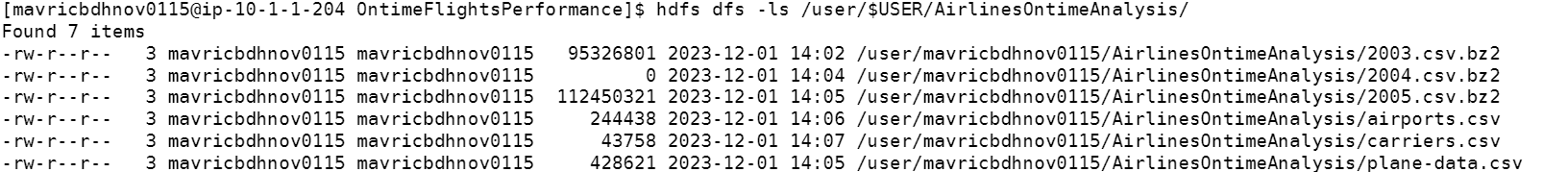
Query 1: Copy the provided csv files in a directory named AirlinesOntimeAnalysis



A black and white image of a number

Description automatically generated





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

rawdata/airline/flights

rawdata/airline/planeinfo

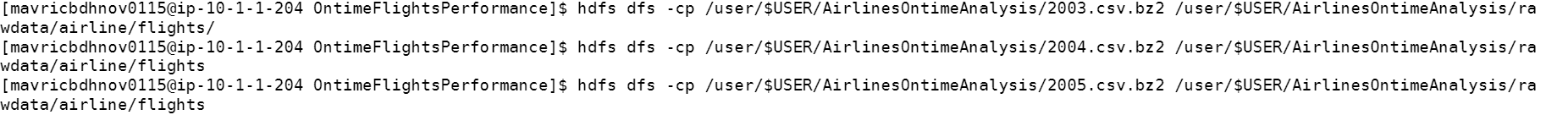
rawdata/airline/airport

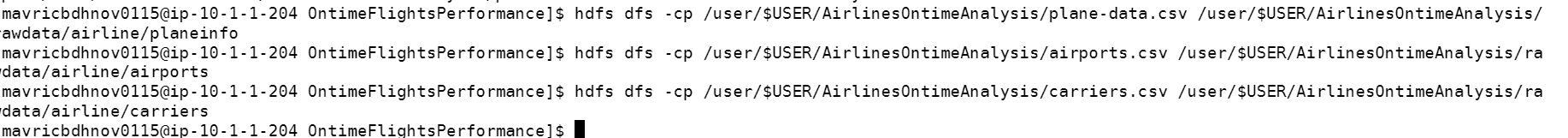
rawdata/airline/carriers

A screenshot of a computer

Description automatically generated

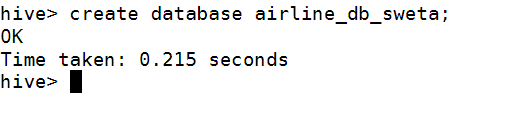
Step 4: **copy the respective files to the above created hdfs structure**





**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Problem-2\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**CREATE DATABASE airline\_db\_sweta;**



1. External airport table created below

CREATE EXTERNAL TABLE airline\_db\_sweta.airports\_stg (

IATA VARCHAR(3),

Name VARCHAR(255),

City VARCHAR(255),

State VARCHAR(255),

Country VARCHAR(255),

Lat DECIMAL(9,6),

Long DECIMAL(9,6)

)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

STORED As TextFile

LOCATION '/user/mavricbdhnov0115/AirlinesOntimeAnalysis/rawdata/airline/airport'

tblproperties(

"skip.header.line.count"="1"

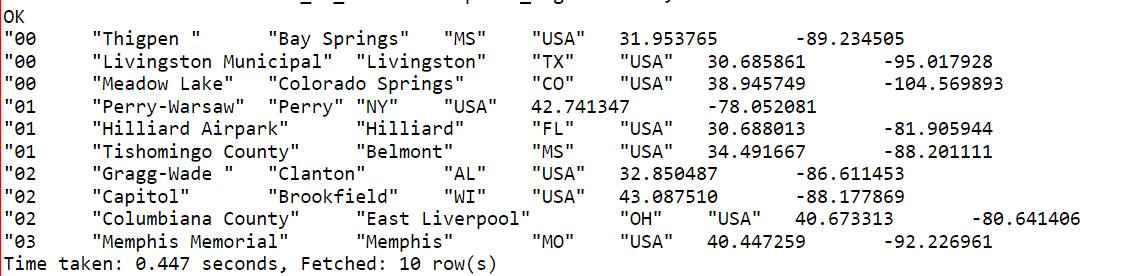
);

A screenshot of a computer

Description automatically generated

LOAD DATA LOCAL INPATH ‘home/mavricbdhnov0115/OneDrive\_1\_11-29-2023/Big Data DV Batch – Assignments/Datasets/OntimeFlightPerformance/airports.csv’ INTO TABLE airline\_db\_sweta.airports\_stg;

Data loaded in airports\_stg



1. External carriers table created below

CREATE EXTERNAL TABLE airline\_db\_sweta.carriers\_stg (

Code VARCHAR(255),

Description VARCHAR(255)

)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

STORED As TextFile

LOCATION '/user/mavricbdhnov0115/AirlinesOntimeAnalysis/rawdata/airline/carriers'

tblproperties(

"skip.header.line.count"="1"

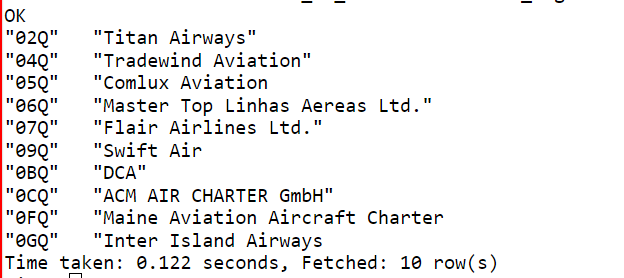
);

A screenshot of a computer code

Description automatically generated

LOAD DATA LOCAL INPATH ‘home/mavricbdhnov0115/OneDrive\_1\_11-29-2023/Big Data DV Batch – Assignments/Datasets/OntimeFlightPerformance/carriers.csv’ INTO TABLE airline\_db\_sweta.carriers\_stg;

Data loaded in carriers table



1. External planeinfo table created below

CREATE EXTERNAL TABLE airline\_db\_sweta.planeinfo (

tailNum VARCHAR(255),

type VARCHAR(255),

manufacturer VARCHAR(255),

issue\_date VARCHAR(10),

model VARCHAR(255),

status VARCHAR(255),

aircraft\_type VARCHAR(255),

engine\_type VARCHAR(255),

year INT

)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

STORED As TextFile

LOCATION '/user/mavricbdhnov0115/AirlinesOntimeAnalysis/rawdata/airline/planeinfo'

tblproperties(

"skip.header.line.count"="1"

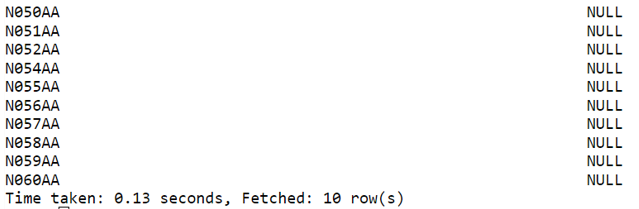
);

A screenshot of a computer code

Description automatically generated

LOAD DATA LOCAL INPATH ‘home/mavricbdhnov0115/OneDrive\_1\_11-29-2023/Big Data DV Batch – Assignments/Datasets/OntimeFlightPerformance/plane-data.csv’ INTO TABLE airline\_db\_sweta.planeinfo\_stg;

Data loaded in planeinfo table



1. External flight table create below

CREATE EXTERNAL TABLE flights (

Year INT,

Month INT,

DayofMonth INT,

DayOfWeek INT,

DepTime INT,

CRSDepTime INT,

ArrTime INT,

CRSArrTime INT,

UniqueCarrier VARCHAR(255),

FlightNum VARCHAR(255),

TailNum VARCHAR(255),

ActualElapsedTime INT,

CRSElapsedTime INT,

AirTime INT,

ArrDelay INT,

DepDelay INT,

Origin VARCHAR(255),

Dest VARCHAR(255),

Distance INT,

TaxiIn INT,

TaxiOut INT,

Cancelled INT,

CancellationCode VARCHAR(1),

Diverted INT,

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/mavricbdhnov0115/AirlinesOntimeAnalysis/rawdata/airline/flights'

tblproperties(

"skip.header.line.count"="1"

);

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

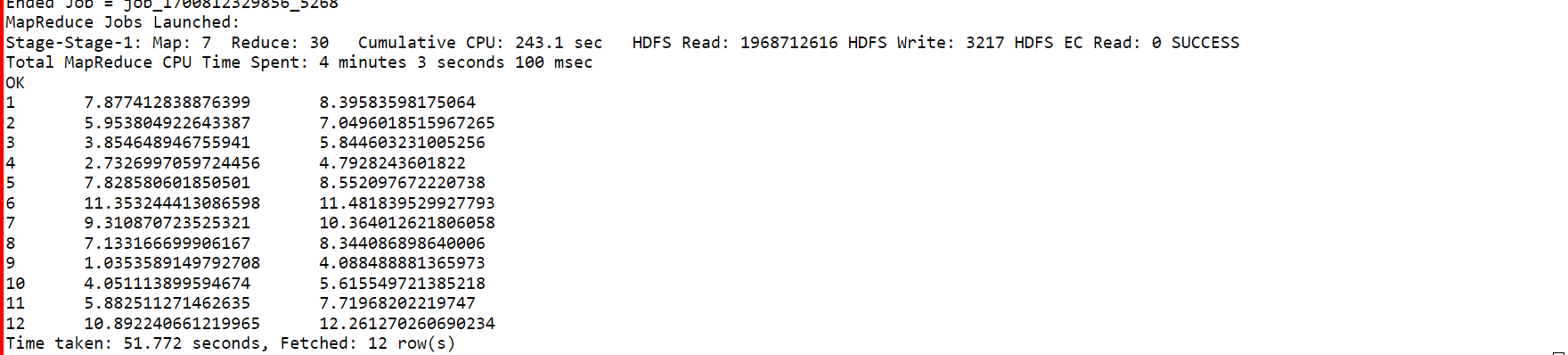
LOAD DATA LOCAL INPATH ‘home/mavricbdhnov0115/OneDrive\_1\_11-29-2023/Big Data DV Batch – Assignments/Datasets/OntimeFlightPerformance/2003.csv’ INTO TABLE airline\_db\_sweta.flights\_stg;

LOAD DATA LOCAL INPATH ‘home/mavricbdhnov0115/OneDrive\_1\_11-29-2023/Big Data DV Batch – Assignments/Datasets/OntimeFlightPerformance/2004.csv’ INTO TABLE airline\_db\_sweta.flights\_stg;

LOAD DATA LOCAL INPATH ‘home/mavricbdhnov0115/OneDrive\_1\_11-29-2023/Big Data DV Batch – Assignments/Datasets/OntimeFlightPerformance/2005.csv’ INTO TABLE airline\_db\_sweta.flights\_stg;

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

Answer: SELECT Month, AVG(ArrDelay) AS AvgArrivalDelay, AVG(DepDelay) AS AvgDepartureDelay FROM airline\_db\_sweta.flights\_stg WHERE Year = 2004 GROUP BY Month;



Question 3: 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

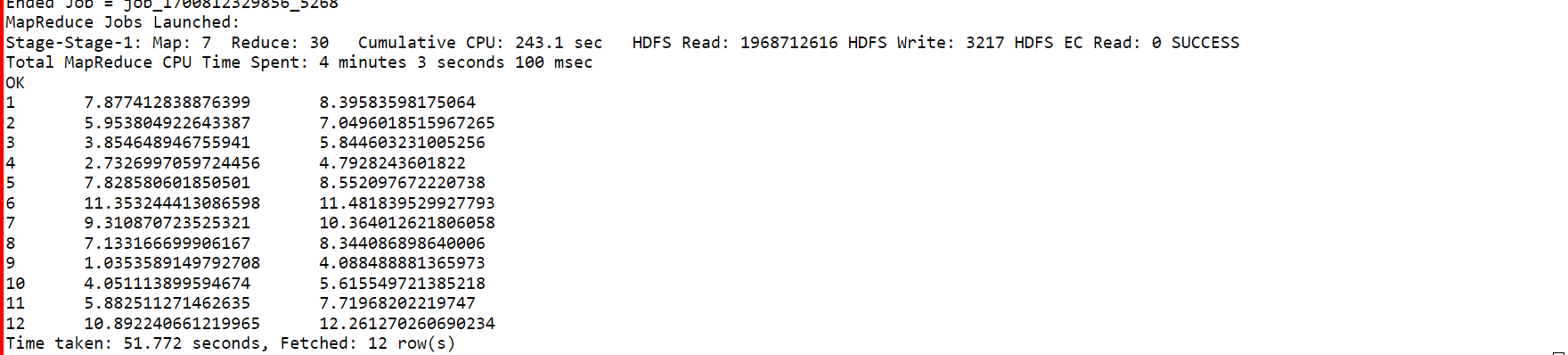
CREATE EXTERNAL TABLE airline\_db\_sweta.flights\_pq (  
    Year INT,  
    Month INT,  
    DayofMonth INT,  
    DayOfWeek INT,  
    DepTime INT,  
    CRSDepTime INT,  
    ArrTime INT,  
    CRSArrTime INT,  
    UniqueCarrier VARCHAR(255),  
    FlightNum VARCHAR(255),  
    TailNum VARCHAR(255),  
    ActualElapsedTime INT,  
    CRSElapsedTime INT,  
    AirTime INT,  
    ArrDelay INT,  
    DepDelay INT,  
    Origin VARCHAR(255),  
    Dest VARCHAR(255),  
    Distance INT,  
    TaxiIn INT,  
    TaxiOut INT,  
    Cancelled INT,  
    CancellationCode VARCHAR(1),  
    Diverted INT,  
    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/mavricbdhnov0115/AirlinesOntimeAnalysis/rawdata/airline/pq\_flights'  
tblproperties(  
"skip.header.line.count"="1"  
);

Loading data into the table flights\_pq

LOAD DATA LOCAL INPATH ‘home/mavricbdhnov0115/OneDrive\_1\_11-29-2023/Big Data DV Batch – Assignments/Datasets/OntimeFlightPerformance/2004.csv’ INTO TABLE airline\_db\_sweta.flights\_pq;

Query : SELECT Month, AVG(ArrDelay) AS AvgArrivalDelay, AVG(DepDelay) AS AvgDepartureDelay FROM airline\_db\_shubham.flights\_stg WHERE Year = 2004 GROUP BY Month;



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

Answer: Yes there is difference of **22** sec

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table type** | **Table Name** | **Start time** | **End time** | **time elapsed** |
| Normal table | airline\_db\_sweta.flights\_stg | Fri Nov 30 15:15:32 +0550 2023 | Fri Nov 30 15:16:29 +0550 2023 | 57sec |
| Parquet Table | airline\_db\_sweta.flights\_pq | Fri Nov 30 15:58:50 +0550 2023 | Fri Nov 30 15:59:25 +0550 2023 | 35sec |