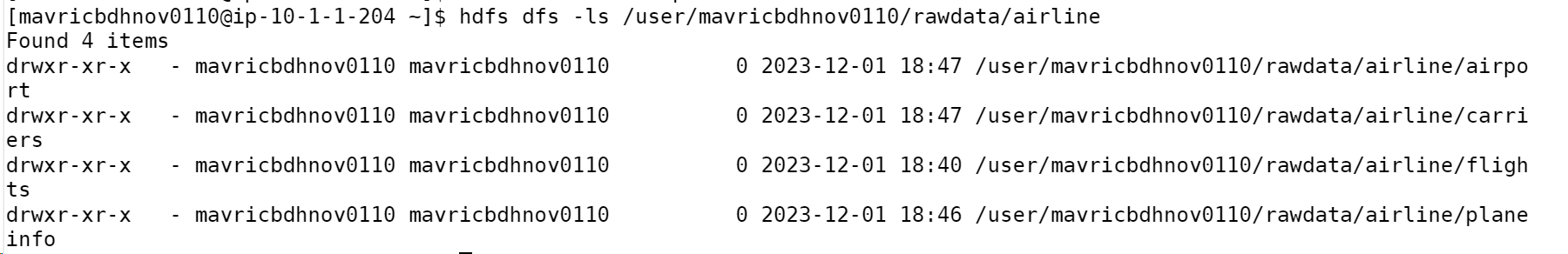
hdfs dfs -mkdir -p /user/mavricbdhnov0110/rawdata/airline/flights

hdfs dfs -mkdir -p /user/mavricbdhnov0110/rawdata/airline/planeinfo

hdfs dfs -mkdir -p /user/mavricbdhnov0110/rawdata/airline/airport

hdfs dfs -mkdir -p /user/mavricbdhnov0110/rawdata/airline/carriers



CREATE DATABASE airline\_db\_shabbir;

USE airline\_db\_shabbir;

A close up of a text

Description automatically generated

CREATE EXTERNAL TABLE Airports\_stg (

    IATA STRING,

    Name STRING,

    City STRING,

    State STRING,

    Country STRING,

    Lat DECIMAL(9, 6),

    Long DECIMAL(9, 6)

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE

LOCATION '/user/mavricbdhnov0110/rawdata/airline/airports'

tblproperties(

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

);

CREATE EXTERNAL TABLE Carrier\_stg (

    Code STRING,

    Description STRING

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE

LOCATION '/user/mavricbdhnov0110/rawdata/airline/carriers'

tblproperties(

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

);

CREATE EXTERNAL TABLE PlaneInfo\_stg (

    tailNum STRING,

    type STRING,

    manufacturer STRING,

    issue\_date STRING,

    model STRING,

    status STRING,

    aircraft\_type STRING

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE

LOCATION '/user/mavricbdhnov0110/rawdata/airline/planeinfo'

tblproperties(

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

);

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 ','

STORED AS TEXTFILE

LOCATION '/user/mavricbdhnov0110/rawdata/airline/flights'

tblproperties(

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

);

A close-up of a white background

Description automatically generated

A close-up of numbers

Description automatically generated

LOAD DATA INPATH 'AirlinesOntimeAnalysis/airports.csv' OVERWRITE INTO TABLE Airports\_stg;

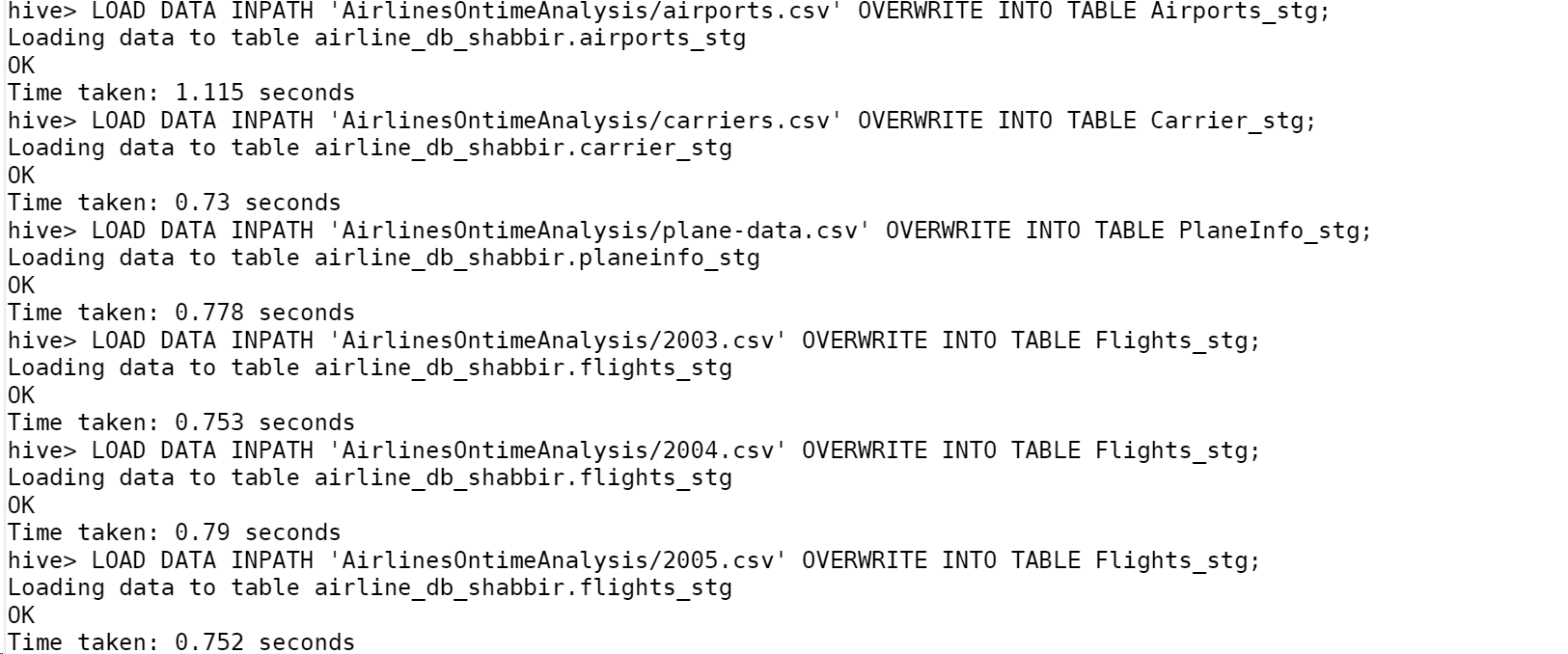
LOAD DATA INPATH 'AirlinesOntimeAnalysis/carriers.csv' OVERWRITE INTO TABLE Carrier\_stg;

LOAD DATA INPATH 'AirlinesOntimeAnalysis/plane-data.csv' OVERWRITE INTO TABLE PlaneInfo\_stg;

LOAD DATA INPATH 'AirlinesOntimeAnalysis/2003.csv' OVERWRITE INTO TABLE Flights\_stg;

LOAD DATA INPATH 'AirlinesOntimeAnalysis/2004.csv' OVERWRITE INTO TABLE Flights\_stg;

LOAD DATA INPATH 'AirlinesOntimeAnalysis/2005.csv' OVERWRITE INTO TABLE Flights\_stg;



A white background with black text

Description automatically generated

Q1.What is the average arrival delay and average departure delay in each

month of the year 2004?

INSERT OVERWRITE DIRECTORY 'workspace\_shabbir/Flight\_delay\_analysis.csv'

SELECT

    Month,

    AVG(ArrDelay) AS avg\_arrival\_delay,

    AVG(DepDelay) AS avg\_departure\_delay

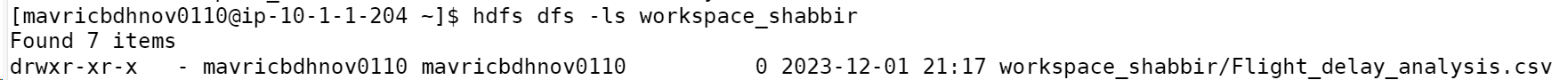
FROM Flights\_stg

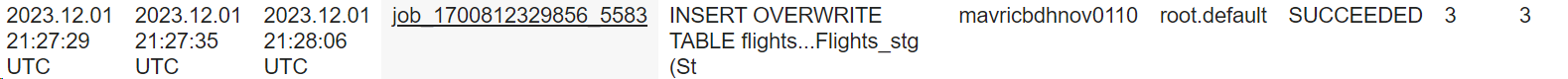
WHERE Year = 2004

GROUP BY Month

ORDER BY Month;

hdfs dfs -cat workspace\_shabbir/Flight\_delay\_analysis.csv





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/mavricbdhnov0110/output/airline/pq\_flights';

A white background with black text

Description automatically generated

A close-up of a computer code

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 avg\_arrival\_delay,

    AVG(DepDelay) AS avg\_departure\_delay

FROM flights\_pq

WHERE Year = 2004

GROUP BY Month

ORDER BY Month;

